GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

EDITED BY

CARL MURCHISON

VOLUME XI

1 9 3 2

Issued monthly by the
CLARK UNIVERSITY PRESS
Worcester, Massachusetts
U. S. A.

$7.00 per volume
Single numbers $2.00

MONTHLY
Two volumes per year

January, 1932
Volume XI, No. 1

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

EDITED BY
CARL MURCHISON

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the post-office at Worcester, Mass., under Act of March 3, 1879

$7.00 per volume
Single numbers $2.00

MONTHLY
Two volumes per year

January, 1932
Volume XI, No. 1

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

GENERAL FACTORS IN TRANSFER OF TRAINING IN THE WHITE RAT*

From the Animal Laboratory of the Department of Psychology, Columbia University

By

THEODORE A. JACKSON

*Accepted for publication by C. J. Warden of the Editorial Board and received in the Editorial Office, May 11, 1931.

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the post-office at Worcester, Mass., under Act of March 3, 1879.

ACKNOWLEDGMENTS

The writer wishes to express his deep appreciation to Professor C. J. Warden of the Department of Psychology, Columbia University, for supervising the experiment throughout, and for much kindly criticism in the preparation of the manuscript. Thanks are also due to Mr. Richard Fitch for many helpful suggestions in the statistical treatment of the results.

THEODORE A. JACKSON

COLUMBIA UNIVERSITY
NEW YORK CITY

CONTENTS

	Acknowledgments	3
I.	Introduction	7
II.	Experiment I	19
III.	Experiment II	39
IV.	Discussion and Summary	49
	References	56
	Résumé en français	58
	Referat auf deutsch	59

I

INTRODUCTION

The presence of transfer of training was noted in the early studies of motor habits in animals, although for the most part its report was incidental to the main purpose of the experiment. Thorndike (18) pointed out that previous experience enables a cat to form associations more quickly; for example, after solving six or eight problem boxes by different methods, the cat develops a tendency to claw at loose objects and does not try to squeeze through holes or to bite at bars. Watson (25) observed that trained rats, those that had been tested on several problems, learn faster than untrained ones. Yerkes (30) also noted, in his work on the dancing mouse, that animals learned a maze more quickly if they had previously run other mazes. Richardson (17) found a distinct difference between the learning of trained and untrained rats, the latter requiring more trials to learn a problem than the former. One of the first investigations of transfer of training in animals was that of Bogardus and Henke (1). They put white rats through a series of mazes, each maze being only slightly different from the preceding one. Transfer was positive in all cases and it was greater when the alteration of pattern was made in the first part of the maze. Hunter (8) compared the learning of untrained pigeons with that of pigeons which had been previously trained in another maze. There was little or no difference between the two groups in the number of trials to learn the maze, although the learning curve

for the trained group showed an earlier drop. That interference may exist is shown by the experiments on squirrels by Yoakum (32), who found that animals which had learned two problem boxes did not learn a third problem box as readily as animals which were put directly on the third problem. The response required for the first two problems was scraping, while for the third it was butting the nose against a latch. Cole (3) has also observed interference in the raccoon when the latch on a problem box was changed in position. Similar observations of interference were made on the monkey by Kinnaman (11) and others.

Somewhat later than the above studies are those of Dashiell (4, 5), Hunter (9), and Ho (7). They all used the white rat and found large positive transfer of training. Dashiell (4), in his earlier study, found that general adaptation to experimental conditions aids the rat in learning a maze, and in a recent study (5) he shows that learning a maze is in part the acquisition of a general orientation toward the food box. Hunter (9) has shown that even between mazes of opposite patterns transfer is positive, and he agrees with Webb in finding the locus of transfer in the first part of the learning curve for the second maze. An experiment by Ho (7) to determine the transfer effect of varying degrees of integration of the first problem gave inconclusive results, although there was some indication that the second habit was more readily acquired the greater the degree of integration of the first habit. Yerkes and Coburn (31) observed that a pig gives evidence of beneficial effects from one problem to another in the multiple-choice type of problem box.

The first systematic work with animals on transfer of training is that of Webb (26). He ran groups of white rats through different pairs of mazes. In order to determine the extent to which transfer depends on the second maze, he ran five groups of animals through the same first maze (Maze A) until they had reached the norm of four out of five errorless trials; then one group was transferred to Maze B, another to Maze C, another to D, another to E, and the last to F. To determine the dependence of transfer on the first maze, the second maze was constant while the first maze varied, as B-A, C-A, D-A, E-A, F-A. In our Table 1 the percentage of savings for the various pairs of mazes is indicated.

The author concluded that (1) the nature of transfer is positive, that is, the learning of one maze has beneficial effects in the mastery of a subsequent maze; (2) that transfer is a composite process consisting of both positive and negative elements, and the total result is determined by the predominance of one or the other of these elements; (3) the degree of transfer is dependent on at least four factors: (*a*) when the first maze is constant the degree of transfer varies with the difficulty of the second maze, (*b*) when the second

TABLE 1

	Trials	Errors	Time		Trials	Errors	Time
A-B	*77.08	85.81	83.77	B-A	80.11	95.22	87.58
A-C	57.85	46.10	34.94	C-A	50.38	62.94	72.67
A-D	69.02	79.71	90.42	D-A	59.14	71.43	74.56
A-E	19.91	54.61	63.40	E-A	42.12	60.03	63.44
A-F	63.01	42.78	59.44	F-A	9.59	51.81	60.46

*Note: 77.08 means that it took 77.08% less trials to learn Maze B after having learned Maze A than it took otherwise.

maze is constant the degree of transfer varies with the difficulty of the first maze, (*c*) the degree of transfer varies with the degree of similarity between the two mazes, and (*d*) the degree of transfer was determined in part by the direction of transfer, that is, if the mazes were not similar there was a greater transfer effect from one maze to the other than was the case when they were learned in the opposite order; (4) the locus of transfer, on the average, was confined to the first five trials, that is, the animals saved the equivalent of the first five trials on the second maze; (5) transfer produced a selective effect on the type of error made—fewer retracing errors were made in the second maze than in the first.

The work of Wiltbank (28) is essentially an extension of that of Webb. The investigation was carried on in the same laboratory and under the same general conditions. The main purpose of his study was to determine whether or not transfer of training was cumulative through a series of mazes. Five different groups of white rats were used, and the order of mazes was rotated in such a manner that each group started with a different maze. His results showed no cumulative effect.

Another problem that he attacked was that concerning the transfer effect of different degrees of partial learning of the first maze on the learning of the second. In one case the initial partial learning was on a less difficult maze, in another it was on a more difficult maze. Table 2 gives the percentages of transfer for the different degrees of partial learning.

TABLE 2

FROM THE LESS DIFFICULT TO THE MORE DIFFICULT MAZE
(MAZE E TO MAZE D)

	After 2 trials in E	After 4 trials in E	After 8 trials in E	After 16 trials in E	After complete mastery of E
Trials	— 3.73	—15.24	— 5.81	34.63	15.44
Errors	—10.60	— 2.65	—39.84	43.50	44.97
Time	75.11	78.76	79.62	72.82	40.13

FROM THE MORE DIFFICULT TO THE LESS DIFFICULT MAZE
(MAZE D TO MAZE E)

	After 2 trials in D	After 4 trials in D	After 8 trials in D	After 16 trials in D	After complete mastery of D
Trials	—22.51	—11.72	— 3.03	47.10	69.22
Errors	—18.55	— 2.17	—44.19	51.01	77.69
Time	75.11	78.95	75.11	71.36	65.41

This part of the study was extended so that each of the above groups, after having completely learned the second maze, was then changed back to complete the learning of the first maze. In this case the results showed positive transfer in nearly all cases. The attempt was also made to determine whether transfer effects were greater from more difficult or less difficult mazes. The results showed slightly greater transfer where the more difficult maze came first. In still another phase of the study an identical part of one maze was inserted in the second maze; however, the results showed no greater saving of errors in the identical part.

In the summary of Wiltbank's results, the following points are made: (1) that transfer between two mazes of the general type used is predominately positive, (2) that transfer through a series of mazes is persistent

although not cumulative, (3) that, when two adjacent mazes have identical parts, the more expeditious learning of the second was not due to a saving of errors in the identical part, (4) that average savings are higher from more difficult mazes to less difficult than the opposite, (5) that the transfer effect between two mazes when the first is only partially learned was not positive until the partial learning was 16 trials, (6) that, when transfer was made from a maze completely learned to one already partly learned, the later maze was learned with a saving in trials when the partial learning had been 2 trials, 4 trials, 8 trials, but transfer was negative for 16 trials.

The foregoing survey of literature on transfer of training in the motor capacities of animals may be summarized as follows:

1. Transfer is predominately positive in mazes of average difficulty.

2. Transfer is not cumulative through a series of mazes.

3. The locus of transfer appears to be in the first part of the learning curve of the second problem.

4. General orientation toward the food box is an important part of maze learning.

5. The similarity between two mazes, as well as their relative difficulty, are important factors in determining the amount of transfer.

6. When two adjacent mazes have identical parts, the more expeditious learning of the second maze does not seem to be due to a saving of errors in the identical part.

7. The degree of transfer varies to some extent with the amount of partial learning on the first problem.

It will be noted that there has been little or no attempt to make an experimental analysis of general factors in transfer of training in animal maze learning. The main purpose of the present study is to deal with certain aspects of this general problem. The fact that transfer has been found to be located principally in the first part of the learning curve of the second problem suggests that it may be due to emotional factors rather than to knowledge elements. That transfer effects are not cumulative through a series of mazes also carries the same suggestion. The fact that transfer effects are greater when the amount of partial learning is greater is also consistent with the idea that transfer may take place through emotional factors rather than knowledge factors. If transfer is produced in such fashion, then initial practice on almost any type of apparatus should result in some degree of positive transfer to the maze. To test this possibility is the purpose of Experiment I of the present study.

Definite conclusions have not as yet been reached concerning the transfer effects from varying degrees of partial learning of the first problem. However, if that initial practice was carried beyond complete learning over into varying degrees of overlearning, the influence of different amounts of initial practice might be made clearer. It is the aim of Experiment II to deal with this aspect of transfer and, if possible, to relate the results with the previous findings on initial partial learning.

In making the present experimental analysis of transfer of training in the white rat the maze was chosen as the principal apparatus to be used. The reasons for this choice were, first, the maze is a convenient apparatus to use with rats as subjects; secondly, the particular maze used, the Warner-Warden design (24), allows of systematic variations in pattern; and, thirdly, because the maze, if not too simple in pattern, is a sufficiently reliable test or measuring device for group differences. Two different patterns, Maze X and Maze Y, as shown in Figures 1 and 2, were used in this investigation. Both patterns contain the same number of blind alleys, and the same number of pathway units. The pathway of Maze X is a "simple zigzag," whereas that of Maze Y is somewhat more complicated and may be designated as a "double zigzag" pattern.

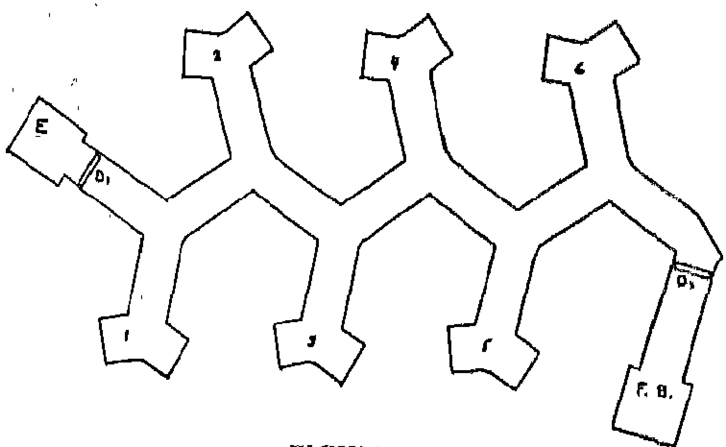

FIGURE 1

PATTERN OF MAZE X

E refers to entrance; *FB* to the food box; *D* to doors; *1, 2, 3, 4, 5,* and *6* to the various blind alleys.

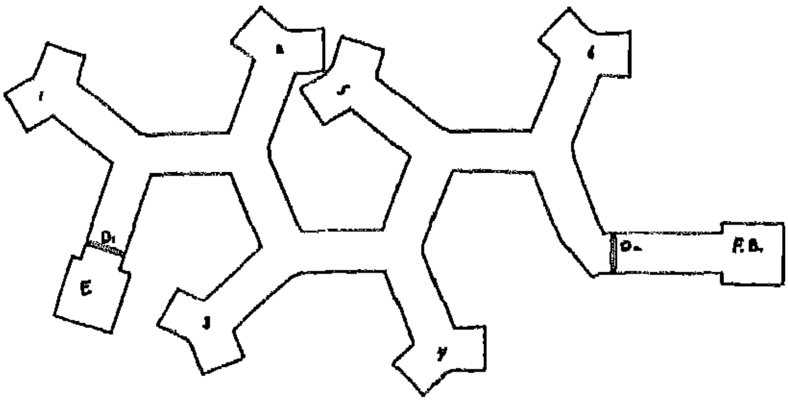

FIGURE 2
PATTERN OF MAZE Y
E refers to entrance; *FB* to the food box; *D* to doors; *1*, *2*, *3*, *4*, *5*, and *6* to the various blind alleys.

While a particular pattern was being used, it was placed on a large 6½- by 8-foot table which was covered with battleship linoleum to provide a seamless floor for the maze. The small room, in which the maze is the only apparatus, was illuminated by a single lamp with a large milk-glass globe which gave quite a diffuse light and was directly above the maze. The only window in the room was darkened by means of a cased-in black window shade, and the doors to the maze room were kept closed during experimentation. In certain of the various transfer situations the following apparatus was used: (*a*) revolving-wheel activity cage (Richter type), (*b*) the Jenkins-Warden motivation apparatus, and (*c*) a simple problem box which the animal operates by stepping on a plate in the floor.

The animals, supplied by the Albino Supply, Inc., Philadelphia, Pa., were medium size, 80 to 100 grams,

and from 60 to 90 days old. Only males were used in order to avoid any possible complications from the oestrous cycle in the female. These animals were housed in small cages, 18 by 12 by 12 inches, which were closed on three sides and had a solid bottom, sawdust being used for bedding. Fresh water was supplied at all times by means of inverted siphon bottles. All animals were fed for one minute in the food box and eight minutes in the feeding cage after their daily run in the maze. The diet was wholewheat bread soaked in milk, supplemented with a weekly ration of greens.

As soon as the animals were received in the laboratory they were taken one at a time and placed in the living cages in random order, allowing five animals to a cage. On the second day their ears were clipped for marking, and on the fourth day they were started on the training of one group or another, see Experiment I. The reason for this three-day period between arrival of the rats and the beginning of the experiment was to allow them to recover from any ill effect of shipment and to become adjusted to the new diet. A longer period could not have been allowed without interfering with the plan of the experiment.

The motivation used throughout was 24-hours' starvation, the incentive employed in the apparatus being a sample of the regular diet. One trial was given a day, and at the end of the trial each animal was allowed to feed one minute in the food-box compartment of the maze, after which it was put back in the living cage. All trials were given in the afternoon between 1:00 and 5:00 P.M. The following records were kept for

each trial: (*a*) time scores, the time required for the animal to go from the entrance to the food box, (*b*) three types of error scores: "A" errors, or entering a blind alley when the animal was oriented toward the food box, "B" errors, or entering a blind alley when the animal was oriented toward the entrance compartment, and "C" errors, or retracing sections of the pathway. A blind-alley error was counted when the animal entered far enough so that the tips of its ears could be seen within. If an animal had not reached the food box in five minutes after it had been inserted in the maze it was removed from the pathway or blind alley and placed in the food-box compartment where it was allowed the regular one-minute feeding period. The norm of mastery used throughout the present study was four errorless trials out of five. The total number of trials to learn was the score used in making the main analysis of results. This measure does not include the first trial, since the latter was considered to be a part of the general preliminary. The five trials of the norm were omitted from the score, as is usual in maze comparisons.

The attempt was made to handle all animals as uniformly as possible. The method finally adopted was to grasp an animal around the body with the left hand in such a way that the thumb and forefinger are at either side of the animal's head. Using such a technique, an animal may be carried to the apparatus and inserted into the pathway with the least disturbance; furthermore, the right hand is free to open and close doors and to operate the stop-watch. During a trial

the experimenter remained very quiet, thus eliminating possible distractions from this source. Additional points of procedure will be mentioned in connection with the different experiments as they are discussed.

II

EXPERIMENT I

The object of this experiment was to determine the extent to which transfer of training in maze learning of the white rat may be due to general factors, such as may be involved in other than maze situations. In order to make such a determination, several groups of animals were tested on the same maze after having previously had different kinds of activity. The various transfer situations used were graduated from very general types of adjustment, such as general laboratory experience, to training on different kinds of apparatus. If such general factors are significant in maze learning, it is important to find out just how much of such activity is necessary to bring about a reliable difference on the maze test, and how much transfer can be brought about by these factors. The first five groups, as shown in Table 3, were given various kinds of general activity previous to the maze test. Group 1 had a minimum of such activity, and each succeeding group was given more and more. The animals in each of these groups were placed singly in new situations, where they were allowed to explore and become adjusted. No food was given, and no particular kind of response was required of them. They were merely given opportunity to get used to a new situation. The last four groups of Table 3 were trained on one or more kinds of apparatus, being rewarded with food after having made a particular response. The groups are arranged in the table in a

TABLE 3

Showing Groups Used

Group number	Number of animals	Activity previous to maze test	Maze test
1	20	None (started directly in Maze X)	Maze X
2	20	3-day maze preliminary (5 min. in entrance and food box of Maze X)	Maze X
3	21	2-weeks' laboratory adjustment; 3-day maze preliminary	Maze X
4	19	2-weeks' laboratory adjustment; 9-day maze preliminary	Maze X
5	20	2 weeks of daily handling; 3-day maze preliminary	Maze X
6	16	2-weeks' daily contact with maze sections; 3-day maze preliminary	Maze X
7	19	2 weeks on problem box, a trial daily; 3-day maze preliminary	Maze X
8	18	*Contact with 3 apparatuses during 2 weeks; 3-day maze preliminary	Maze X
9	18	2-weeks' laboratory adjustment; 3-day maze preliminary; learned Maze Y	Maze X

*First four days, revolving wheel; next five days, Jenkins-Warden motivation apparatus; last five days, simple problem box.

progressive series according to the total amount of activity in the transfer situation.

Group 1 was put directly into the maze after being kept in the laboratory three days. Group 2, in addition to this, was given the usual three days of preliminary activity previous to the maze test. This preliminary consisted of a five-minute feeding period in the entrance compartment and another in the food-fox compartment daily. During this procedure the animal did not have access to the pathway of the maze, but was

allowed to explore over the top of the apparatus near the compartment in which it was feeding. The transfer situation for Group 3 consisted of a two-weeks' general laboratory adjustment followed by a three-day maze preliminary. They were fed regularly, just as in the case of animals being used on maze tests, but were handled as little as possible. Group 4 was given a similar two-weeks' laboratory adjustment, but the maze preliminary was increased to nine days instead of three. The transfer situation for Group 5 involved two weeks of handling which corresponded to the period of laboratory adjustment of the two previous groups, this being followed by the regular three-day maze preliminary. The handling in this case was designed to be approximately equivalent to the amount involved when an animal was being used in an apparatus. The animals were carried from their living cages and put in other living cages which were located at approximately the same distance away as the maze.

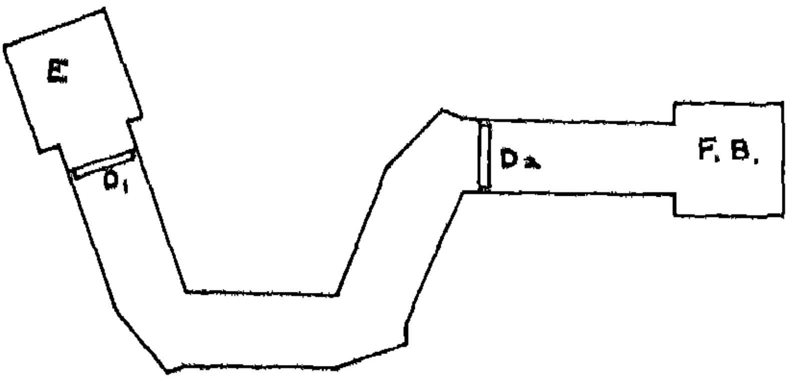

FIGURE 3
Arrangement of Maze Sections Used with Group 6

After five minutes' exploration in one cage each animal was placed in another cage for a similar period, then put back in the feeding cage and fed, and finally back into the living cage again. This procedure enabled the experimenter to keep constant the amount of handling from day to day and from rat to rat. Group 6 was put in a transfer situation which allowed it to make contact with maze materials as shown in Figure 3. Two weeks of daily runs were given through the three pathway sections of the maze. This activity constituted a kind of training, since each animal was rewarded with a sample of the regular diet in the food compartment box at the end of the run. The transfer situation for Group 7 consisted in two weeks of daily trials in a problem box, and maze preliminary. Group 8 was given training on three kinds of apparatus as follows: (*a*) four days in a revolving-wheel activity cage, (*b*) five days in the Jenkins-Warden motivation apparatus, and (*c*) five days in the problem box. The animal was put in the revolving wheel for four minutes each day; in the Jenkins-Warden motivation apparatus each animal was allowed to run across the grid, without shock, to food, being allowed only a nibble, and was then put back in the entrance compartment for another run. This procedure was kept up for three minutes, after which one minute of uninterrupted feeding was allowed in the food compartment. The practice on the problem box was given just as in the case of Group 7, with the usual three-day preliminary. Group 9 had considerably more training than the other groups. It was given the two-weeks' laboratory ad-

justment, followed by a maze preliminary on Maze Y, then, after learning this maze, was tested on Maze X. An average of 23 trials was required to learn Maze Y. In comparing this group with any others, it must be kept in mind that the transfer situation in this case was considerably longer than the two-weeks' period involved in the case of the other groups.

TABLE 4

DISTRIBUTION OF SCORES, IN TRIALS TO LEARN, FOR THE VARIOUS GROUPS IN THE LEARNING OF MAZE X

Number of trials to learn maze	Groups 1	2	3	4	5	6	7	8	9
0				1	1		1	2	
1		1			2			1	3
2	2	1	3	2	4	3	7	3	4
3	2	2	6	2	4	3	3	4	3
4	2	2	3	2		4	3	3	2
5	2	2	3	4	3	2	5	1	1
6	1	3	1	4	1	2		2	4
7	3	3		2	3	1		1	1
8	2	3	2		2	1			
9	2	1	3	1				1	
10	1	1		1					
11	1	1							
12	1								
13									
14	1								

TABLE 5

SHOWING AVERAGE TRIALS TO LEARN AND STANDARD DEVIATIONS FOR THE VARIOUS GROUPS

Group number	Number of animals	Average	Standard deviation	*S.D.* of the average
1	20	6.80	3.30	.74
2	20	6.00	2.59	.58
3	21	4.76	2.39	.52
4	19	5.00	2.36	.54
5	20	4.00	2.45	.55
6	16	4.25	1.75	.44
7	19	3.16	1.42	.33
8	18	3.56	2.36	.55
9	18	3.56	1.94	.46

The results of this experiment show definitely that there is positive transfer of training from the various types of transfer situations used to maze training. Table 4 shows the distribution of scores, in trials to learn, for the various groups. It is evident from inspection that the range of scores is smaller for those groups which had the greater amount of pre-maze activity. The arithmetic means and standard deviations for the groups are shown in Table 5 and Figure 4. The smaller average number of trials to learn in the case of the groups with greater previous experience clearly indicates the transfer effect. The correlation coefficient of —.88 between the amount of pre-maze activity and the number of trials required to learn indicates the same trend.

The reliability of the differences between group

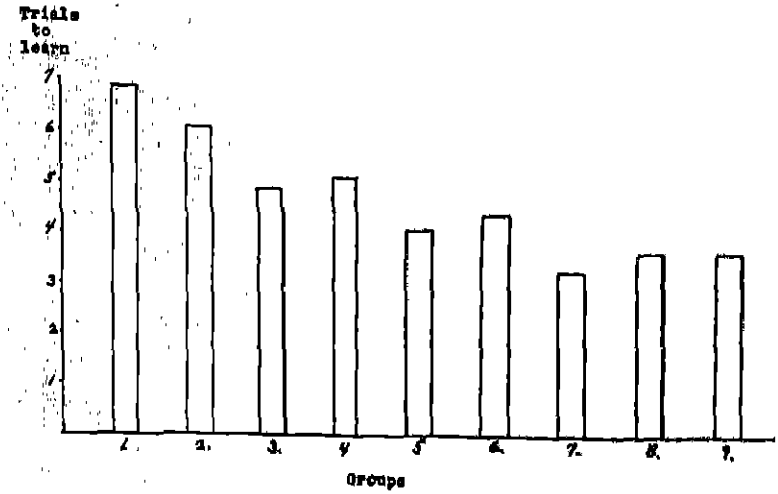

FIGURE 4
Showing Average Number of Trials to Learn Maze X

TABLE 6

Reliability of the Differences between Averages of Various Groups

Groups	Difference	S.D. of the difference (obtained)	*S.D. of the difference (true)	Difference / S.D. of the difference (D/SDD)	Chances in 100 of a true difference greater than zero†
1 & 2	.80	.94	.82	.98	83.65
1 & 3	2.04	.90	.79	2.58	99.51
1 & 4	1.80	.92	.80	2.25	98.78
1 & 5	2.80	.92	.80	3.50	100.00
1 & 6	2.55	.86	.75	3.40	100.00
1 & 7	3.64	.81	.71	5.13	100.00
1 & 8	3.24	.92	.80	4.05	100.00
1 & 9	3.24	.87	.76	4.26	100.00
2 & 3	1.24	.78	.68	1.82	96.56
2 & 4	1.00	.79	.69	1.45	92.65
2 & 5	2.00	.80	.70	2.86	99.79
2 & 6	1.75	.73	.64	2.73	99.68
2 & 7	2.84	.66	.58	4.90	100.00
2 & 8	2.44	.80	.70	3.49	100.00
2 & 9	2.44	.74	.65	3.75	100.00
3 & 4‡	— .24	.75	.65	— .37	65.43
3 & 5	.76	.76	.66	1.15	87.49
3 & 6	.51	.67	.58	.88	81.06
3 & 7	1.60	.62	.54	2.96	99.85
3 & 8	1.20	.75	.65	1.85	96.78
3 & 9	1.20	.69	.60	2.00	97.72
4 & 5	1.00	.76	.66	1.52	93.57
4 & 6	.75	.70	.61	1.23	89.07
4 & 7	1.84	.62	.54	3.41	100.00
4 & 8	1.44	.77	.67	2.15	98.42
4 & 9	1.44	.71	.62	2.32	98.98
5 & 6	— .25	.70	.61	— .41	65.91
5 & 7	.84	.64	.57	1.47	92.92
5 & 8	.44	.78	.68	.65	74.22
5 & 9	.44	.71	.62	.71	76.11
6 & 7	1.09	.55	.48	2.27	98.84
6 & 8	.69	.71	.62	1.11	86.65
6 & 9	.69	.63	.55	1.25	89.44
7 & 8	— .40	.64	.57	— .71	76.11
7 & 9	— .40	.56	.49	— .82	79.39
8 & 9	.00	.72	.63	.00	0.00

*See Section 4 for computation of true *S.D.* of the difference.

†See McCall for criterion of a true difference; *D/SDD* should be 2.78 or greater.

‡A minus sign is placed before the difference and before *D/SDD* when the first group of the two compared has a smaller average.

averages is shown in Table 6. It may be noted that there is a reliable difference between Groups I and 5. This means that the group which had been merely handled made a reliably lower average score when tested on the maze than the untrained group. All other groups which had had more experience than the amount involved in handling also showed reliable differences when compared with an untrained group. The differences between Groups 1 and 3, and 1 and 4, were also almost reliable, the D/SDD being 2.58 and 2.25, respectively. There was no greater transfer for Group 4 than for Group 3, in spite of the fact that the former had six extra days of preliminary activity. It is evident, therefore, that the latter is ineffective in inducing greater transfer effects. This seemed rather strange since the three-day preliminary itself exerted considerable influence. Another unexpected result is that between Groups 5 and 6. Both groups had received the same amount of handling, but Group 6 had also made contact with maze materials during its pre-maze activity. Still, the greater amount of transfer occurred in the case of Group 5. Evidently mere running through maze sections without blind alleys has little or no transfer significance. In the case of Groups 7 and 8, the latter had had more experience, since it had been trained on two more kinds of apparatuses. But the former showed the greater transfer effect. This may seem a bit surprising, but, since the difference is not reliable, it may very well be due to chance. At least greater transfer to the maze was not effected by training on three apparatuses rather than

one, if the total amount of practice is constant. The most unexpected result appears in the comparison of the average score of Group 9 with other averages. A somewhat less amount of transfer occurred from maze to maze than took place from problem box to maze.

As will be seen by referring to Tables 3 and 5, the transfer situations used may be thrown into four more general groups on the basis of similarity. Combinations were made as follows: Groups 1 and 2, *Untrained;* Groups 3 and 4, *Adjusted;* Groups 5 and 6, *Handled*; Groups 7 and 8, *Trained;* and, of course, Group 9 which was *Maze-trained* stands alone. The formation of these combinations is not only justified on the basis of similarity of transfer situations but also because the reliability of the difference between any two groups going into the same combination is low, being less than one in all cases. In the computation of averages of the combined groups, weight was given to the single group averages according to their respective populations. And the computation of standard deviations was by a formula found in Yule (33, p. 142).

The arithmetic means and the standard deviations for the various combinations are shown in Table 7. Here,

TABLE 7

AVERAGES AND STANDARD DEVIATIONS OF TRIALS TO LEARN FOR THE COMBINED GROUPS

Name of combination	Groups in combination	Number of animals	Average	Standard deviation	S.D. of the average
Untrained	1 & 2	40	6.40	2.99	.47
Adjusted	3 & 4	40	4.87	2.38	.38
Handled	5 & 6	36	4.11	1.74	.29
General training	7 & 8	37	3.35	1.94	.32
Maze-trained	9	18	3.56	1.94	.46

TABLE 8

RELIABILITY OF THE DIFFERENCE BETWEEN AVERAGES OF TRIALS TO LEARN FOR THE COMBINED GROUPS

Combinations compared	Difference	S.D. of the difference (obtained)	S.D. of the difference (true)	Difference / S.D. of the difference (D/SDD_t)	Chances in 100 of a true difference greater than zero
Untrained & Adjusted	1.53	.60	.53	2.89	100
Untrained & Handled	2.29	.55	.48	4.77	100
Untrained & General Training	3.05	.57	.50	6.10	100
Untrained & Maze-trained	2.84	.66	.57	4.99	100
Adjusted & Handled	.76	.48	.42	1.81	96.49
Adjusted & General Training	1.52	.50	.44	3.45	100
Adjusted & Maze-trained	1.31	.60	.52	2.52	99.41
Handled & General Training	.76	.43	.38	2.00	97.72
Handled & Maze-trained	.56	.54	.47	1.19	88.30
General Training & Maze-trained	—.21	.56	.49	— .43	66.64

as before, the combination with the more previous experience makes lower average scores on the maze test. The reliability of the differences between different combinations is shown in Table 8. It may be seen that there is a reliable difference between the average score made by the untrained group and any other group, although the reliability indices are higher for the groups

with the greater amount of experience. The fact of a reliable difference between the untrained and the adjusted may be interpreted to mean that those animals which had had a two-weeks' laboratory adjustment and a three-day maze preliminary make reliably lower scores when tested later on the maze than those which had not had such experience. The relative importance of the remaining transfer situations may be seen by examining the reliability of the difference between the corresponding group averages. Since the handled group had had all the experience of the adjusted group, in addition to being handled, a rough indication of the significance of handling *per se* may be had by comparing the adjusted and the handled groups (see Table 8). Similarly, an indication of the importance of general training *per se* is obtained by comparing the handled with the group which had general training. All the factors may thus be measured by the usual method employed in computing residual transfer effects. When this has been done, the relative importance of the several factors is as follows: adjustment, 2.89; handling, 1.81; general training, 2.00; maze training, 1.19.

TABLE 9

SHOWING AMOUNT AND PERCENTAGE OF SAVINGS IN TRIALS TO LEARN THE MAZE AND IN AVERAGE TOTAL ERRORS FOR THE VARIOUS COMBINED GROUPS

Combined group	Savings in trials to learn		Savings in total errors	
	Amount	*Percentage*	*Amount*	*Percentage*
Adjusted	1.5	23.9	4.8	33.1
Handled	2.3	35.8	5.3	36.6
General trained	3.1	47.7	8.2	56.6
Maze-trained	2.8	44.4	8.2	56.6

The savings in trials to learn and in total "A" errors, appear in Table 9, and show the same general trend concerning transfer effects. The greater the previous experience, the lower the score when tested in the maze. It may be noted that the two sets of scores are very closely related, the adjusted group having the smallest amount of transfer in both cases, and the group with general training evidencing the greatest amount of transfer effect. The saving in the case of the handled group is almost the same in both scores, and ranks in either case between the adjusted and the maze-trained groups. Although the group with general training shows more saving than the maze-trained group when measured by the number of trials to learn the maze, the two groups show equal saving in total error scores. This difference is made clear when the error-learning curve of the two groups is examined (see Figure 5).

It may be noted that the maze-trained group effected a greater saving of errors on the first two or three trials, and this resulted in a relatively lower total error score.

In general, adjustment to laboratory conditions, etc., was markedly effective in producing transfer to maze learning, and handling was somewhat more effective. General training in other than maze apparatus is still more effective in bringing about transfer to the maze, and such transfer effect is apparently as great as that between two mazes, even when the maze-training period was longer.

It is obvious that there can be no transfer of specific knowledge elements in the case of the adjusted and the handled groups where no apparatus was used in the

training situations. It seems clear, also, that the same can be said of the groups given general training, since the types of apparatus used were in no way similar to the test apparatus. In fact, only in the case of Group 9, which was transferred from maze to maze, does there appear to be any possibility of specific identical elements. Here the transfer was less than from an entirely diverse apparatus, which suggests that specific knowledge elements may not have been of much importance even in this case. It thus appears that general factors of some sort must play a major rôle in the usual transfer experiment on the maze. While the nature of these general factors cannot be precisely determined from the data at hand, some suggestions may be in order regarding the factors that might have been operative under the several conditions—adjustment, handling, and general training.

Adjustment, which consisted mainly of two weeks spent in the laboratory, allowed the animals to recover from any ill effects of shipping, such as starvation, unfavorable temperature conditions, etc. During this time the fact that they became accustomed to living in metal cages may have helped them to adjust later to the maze which was also made of metal. Getting used to the new diet may also have been a factor, and the adjustment to the rhythm of feeding may have been of equal importance. The three-day preliminary on the test maze gave the animals opportunity to explore the external features of the apparatus and become adapted to being placed in a strange situation alone. Feeding in the entrance and food box doubtless caused an asso-

ciation to be formed between the maze and food, and the response of feeding in the maze may have eliminated fear to some extent.

In handling, the animals naturally got used to being picked up by the experimenter and removed to another place. The amount of squirming and wriggling was noticeably reduced in the two-weeks' period. In the transfer situation, handling involved putting each animal singly into two different cages for five minutes each. The factor of exploring these new situations may also have had some transfer value.

General training in an apparatus involved all the experience included in adjustment and in handling, and, in addition, gave opportunity to explore a strange situation which was very unlike either the living cage or the maze. During this training the animals seemed to develop a tendency to react to the food compartment, and, when the animal was transferred to the maze, vigorous exploratory movements began almost immediately; apparently, an association had been formed between apparatus and securing food, and the connection was sufficiently general to carry over from one apparatus to another.

In maze training, all the above-mentioned factors were involved, and, in addition, certain specific knowledge elements. By knowledge elements in the maze is meant specific turns, "blind-alleyness," as contrasted with true pathway conditions, and the like. But specific knowledge may exert either positive or negative transfer effect, as has been shown by Webb (26). Apparently, the net transfer effect from maze to maze

was somewhat less, because, although all the general factors which produce positive transfer were operative, there were, in addition, specific knowledge elements which produced negative transfer.

The question as to the locus of transfer, or the stage in the learning process at which transfer is most effective, has offered an interesting problem to all students of this general topic. The locus of transfer is usually found by comparing the learning curve of a group which has had some sort of previous training with the curve of another group, used on the same problem, which has not had previous training.

As is well recognized, ordinary learning curves show distortion due to the dropping out of the animals as they master the problem. This involves a gradual decrease in population and results in an undue weighting of the scores of the poorer rats. In order to avoid this, the following method was used in plotting the curves of Figures 5 and 6. The score of each animal was extended to the sixteenth trial even though they had completed the learning earlier, because at that point the several curves were approximately flat. Time values in such cases were computed on the basis of four seconds per trial which was the modal score in the trials comprising the norm. These trials in this extension were considered errorless, since the animals had actually mastered the problem. The time curves shown in Figure 6 were also smoothed by computing the average in each case on the middle 80% of the scores. Since the distributions involved are approximately normal, the average is not disturbed by this procedure, and more representative curves are obtained.

In Figure 5, based on error scores taken from Table 10, the learning curves for the various groups may be examined. The curve for the untrained group is the heavier line and is uppermost in all but the first trial. The locus of transfer for any other group will be determined by comparing the curve of that group with the curve for the untrained. The curves for the adjusted and the handled appear to take about the same course, there being considerable criss-crossing in the middle portion of the curve. It may be noted that these curves start at about the same point as the curve for the untrained group, thus, for these groups, transfer was not in evidence on the first trial. However, con-

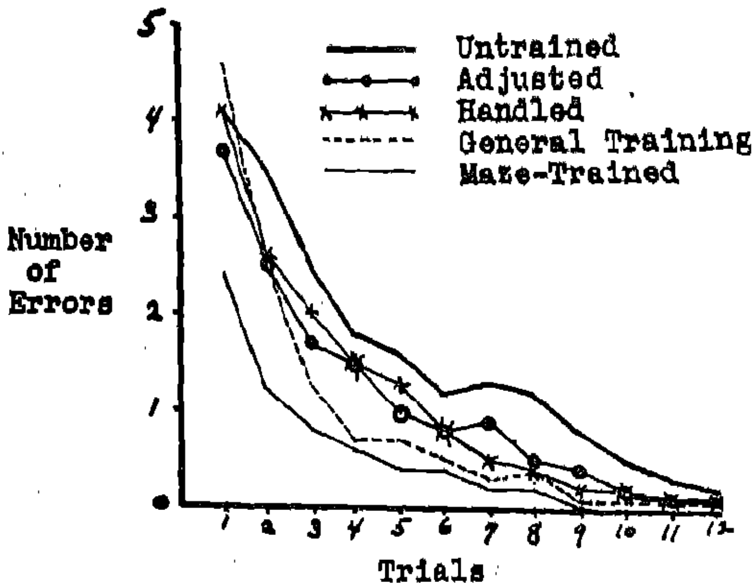

FIGURE 5

Learning Curves Based on Error Scores for the Combined Groups

TABLE 10

AVERAGE "TIME" AND AVERAGE "ERROR" SCORES FOR EACH TRIAL FOR THE VARIOUS COMBINATIONS

Trial number	Untrained Time	Untrained Error	Adjusted Time	Adjusted Error	Handled Time	Handled Error	General training Time	General training Error	Maze-trained Time	Maze-trained Error
1	150.2	4.1	107.9	3.7	121.5	4.1	147.9	4.6	29.1	2.4
2	97.4	3.5	55.9	2.5	48.9	2.6	35.3	2.5	13.4	1.2
3	60.4	2.5	26.8	1.7	18.7	2.0	24.2	1.3	10.9	.8
4	25.5	1.8	14.8	1.5	12.4	1.5	13.8	.7	9.7	.6
5	14.0	1.6	11.3	1.0	9.2	1.3	11.4	.7	9.4	.4
6	17.8	1.2	10.9	.8	7.2	.8	8.4	.5	8.1	.4
7	8.1	1.3	16.2	.9	6.5	.5	10.9	.3	7.2	.2
8	7.0	1.2	11.7	.5	5.6	.4	10.9	.4	6.1	.2
9	5.7	.8	7.6	.4	4.8	.2	5.2	.1	4.6	.0
10	5.5	.5	6.2	.2	4.2	.2	5.4	.1	4.3	.0
11	4.9	.3	5.3	.1	4.4	.1	4.1	.1	4.2	.0
12	4.2	.2	5.0	.1	4.2	.1	4.1	.0	4.0	.0
13	4.5	.2	4.4	.1	4.1	.0	4.1	.0	4.0	.0
14	4.3	.1	4.3	.1	4.0	.0	4.0	.0	4.0	.0
15	4.1	.1	4.1	.0	4.0	.0	4.1	.0	4.0	.0
16	4.2	.1	4.0	.0	4.0	.0	4.0	.0	4.0	.0

2. The following general factors appear to be important in inducing transfer to the maze: (*a*) adjustment, i.e., getting over any ill effects due to shipment, becoming accustomed to the physical environment of the laboratory, adaptation to the new diet and the rhythm of feeding; (*b*) handling, i.e., getting used to being picked up and carried about by the experimenter; and (*c*) general training, i.e., the development of a general tendency to explore an apparatus vigorously, a tendency which is doubtless due to the association of food with the apparatus.

3. The locus of transfer induced by general factors is largely limited to the first five or six trials of the second problem, although much less transfer was apparent on the first trial than occurs in maze-to-maze transfer.

III

EXPERIMENT II

The purpose of this experiment was to determine the transfer effect of different degrees of overlearning of one maze on the learning of another maze. It has already been shown by Wiltbank (28) and others that transfer effects vary somewhat with the degree of learning of the first maze, but no one has yet attempted to find the effects of different degrees of overlearning on transfer.

The two mazes used in this experiment were the same as those described in Experiment I, Maze X being used in the training, and Maze Y later for testing. The animals employed were also the same as in Experiment I, each group in this experiment having been selected at random from Groups 1 to 8 of Experiment I. Table 11 shows the degree of overlearning for each

TABLE 11

SHOWING GROUPS USED IN EXPERIMENT II

Group letter	Number of animals	Amount of practice on Maze X	Standard problem
A	20	None (2-weeks' laboratory adjustment, 3-day preliminary in Maze Y)	Maze Y
B	42	Learned Maze X to norm (4 perfect out of 5)	Maze Y
C	6	10 trials beyond norm in Maze X	Maze Y
D	18	20 trials beyond norm in Maze X	Maze Y
E	24	30 trials beyond norm in Maze X	Maze Y
F	28	40 trials beyond norm in Maze X	Maze Y

TABLE 14

RELIABILITY OF THE DIFFERENCE BETWEEN THE AVERAGES OF THE VARIOUS GROUPS IN LEARNING MAZE Y

Groups compared	Difference	*S.D.* of the difference (obtained)	*S.D.* of the difference (true)	Difference / *S.D.* of the difference (D/SDD_t)	Chances in 100 of a true difference greater than zero
A & B	1.82	3.54	3.45	.53	70.19
A & C	8.25	5.76	5.53	1.49	93.19
A & D	5.86	4.30	4.19	1.40	91.92
A & E	3.50	4.51	4.40	.80	78.81
A & F	4.07	3.67	3.58	1.14	87.29
B & C	10.07	4.99	4.87	2.07	98.08
B & D	7.68	3.33	3.25	2.36	99.09
B & E	1.68	3.63	3.54	.48	68.44
B & F	2.25	2.47	2.41	.93	82.38
C & D	2.39	5.55	5.41	.44	67.00
C & E	11.75	5.79	5.65	2.08	98.12
C & F	12.32	5.08	4.95	2.49	99.36
D & E	9.36	4.45	4.34	2.16	98.46
D & F	9.93	3.47	3.38	2.94	100.00
E & F	.57	3.85	3.75	.15	55.96

TABLE 15

SHOWING AVERAGE TRIALS TO LEARN AND STANDARD DEVIATIONS FOR THE COMBINED GROUPS IN LEARNING MAZE Y

Group letter	Number of animals	Average	Standard deviation	*S.D.* of the average
B	42	20.43	10.46	1.61
C, D	24	28.71	12.22	2.49
E, F	52	18.44	13.37	1.85

ing of the first problem. Transfer was small and positive in the case where the shift was made to the *second problem at zero overlearning*, i.e., just as soon as the norm was reached on the first. For 10 to 20 trials of overlearning the transfer effect was negative, but for 30 to 40 trials the effect was positive again. The reliability indices for the differences between these

groups, as shown in Table 14, are not significant, hence it was decided to combine certain groups. Groups C and D were combined because they were instances of negative transfer and their averages were nearly the same; Groups E and F, having about the same average, although showing positive transfer, were also combined. The average and standard deviation of the combined groups are shown in Table 15, and the reliability of the difference between the averages in Table 16. From the latter table it may be seen that the difference between Groups B and C-D, or between zero overlearning and 10 to 20 trials overlearning, was significant. And the difference between the latter group and Group E-F, which had 30 to 40 trials overlearning, was also significant.

Something of the nature of the factors involved may be shown by an analysis of the persistance of the behavior pattern from the first to the second maze. Since the first maze or transfer situation was identical for the different groups, it might be supposed that the

TABLE 16

RELIABILITY OF THE DIFFERENCE BETWEEN THE AVERAGES OF THE COMBINED GROUPS

Groups compared	Difference	S.D. of the difference (obtained)	S.D. of the difference (true)	Difference / S.D. of the difference (D/SDD$_t$)	Chances in 100 of a true difference greater than zero
B & C, D	8.28	2.96	2.89	2.87	100.00
B & E, F	1.99	2.45	2.39	.83	79.67
C, D, & E, F	10.27	3.10	3.02	3.40	100.00

TABLE 17

SHOWING FREQUENCY AND PERCENTAGE OF "A" ERRORS IN THE DIFFERENT CULS-DE-SAC FOR THE COMBINED GROUPS

Cul-de-sac number	B Freq.	B %	C-D Freq.	C-D %	E-F Freq.	E-F %
1	203	12.7	120	9.4	176	9.9
2	453	28.3	405	31.6	520	29.3
3	82	5.1	115	9.0	134	7.6
4	557	34.8	515	40.2	694	39.2
5	55	3.4	27	2.1	46	2.6
6	252	15.7	100	7.8	202	11.4
Total	1602		1282		1772	

NOTE: The probability that the two distributions are due to the same factors is as follows:

Groups B and C-D	.000015
Groups C-D and E-F	.0156

distribution of errors in the second maze would be the same regardless of the amount of overlearning in the first maze. However, this was not the case as may be seen by reference to Table 17. In fact, a considerable variation in the percentage of errors made in the different alleys occurred under the various degrees of mastery. The difference in type of response made from the several degrees of overlearning can best be shown by applying the Pearson Chi-Square test to the error distributions of Groups B, C-D, and E-F.

The Chi-Square test is a method for determining whether or not two given curves are significantly different. Although, in most applications of the test, one of the two curves has been a theoretical one—such, for example, as the normal curve—it may be used in comparing two empirical curves, as has been done by Wilson (27). The formula used in the latter case is as follows:

$$X^2 = \frac{N_1 N_2}{10{,}000} S \left\{ \frac{(p_1 - p_2)}{(n_1 - n_2)} \right\}$$

where p_1 and p_2 are percentages of entrances into the different blind alleys, n_1 and n_2 are the numbers of such entrances, N_1 and N_2 are the total number of entrances in all alleys, and S is a summation sign. When the formula is applied the result must be converted into a probability value by the use of Pearson's Tables (15). The value there found indicates the probability that the two curves are a result of the same factors. This would mean, of course, that the smaller the index the greater the likelihood that the two curves are produced by different factors.

When application of the method was made to the distribution of errors in the present data, the following results were obtained. The probability that the underlying factors in the learning of Group B were the same as in Group C-D was equal to .000015. The corresponding value between Groups C-D and E-F was .0156.* This seems to mean that the responses made in Maze Y differ significantly between zero and slight overlearning of Maze X. Furthermore, when overlearning of the first maze was greater, there was a tendency for the response made on Y to be somewhat more like the type made after transfer at mastery, since the index was greater in that case. That is, there appears to be the greatest persistance of the first maze habit, and thus the most interference effect, when transfer was made following a slight amount of overlearning. The amount of persistance in that case was greater than when transfer was made at either zero

overlearning or with the greatest amount of overlearning used in the present investigation. A possible explanation of these facts has been suggested by Razran (16), who supposes that there may be stages in complex learning corresponding to the "generalization" and "differentiation" phases of specific conditioning. The "generalization" phase of maze learning would presumably be represented by Group C-D which was transferred somewhat beyond the usual norm of mastery. The "differentiation" phase would then be represented by Group E-F in which the fixation process had been carried much further. Accordingly, if transfer is made at the period of "generalization," there would be the greatest amount of habit interference, or persistence, of the first habit. If transfer follows a greater amount of practice, that is, after the first habit has become "differentiated," there would be less persistance of the original habit in the second maze. The facts at hand seem to correspond approximately to the expectations from this suggestion.

The question now arises concerning the relation between partial learning and overlearning as regards transfer effects. A preliminary study of partial learning was made by Wiltbank and was discussed in the introduction. Although his general results are not strictly comparable to my own, it may be pointed out that the mazes he used were only slightly more difficult than those used in the present experiment. The curves based on the two sets of data have been placed in juxtaposition in Figure 8 so as to bring out the relation between partial and overlearning. When both curves are

considered, the following relationships between degree of mastery and transfer effect seem to hold:

1. Slight partial learning (2-8 trials)	Small negative transfer
2. Medium partial learning (16-20 trials)	Large positive transfer
3. First maze mastered (*norm 4 of 5*)	Small positive transfer
4. Slight overlearning (10-20 trials)	Large negative transfer
5. Large overlearning (30-40 trials)	Medium positive transfer

It may be in order at this point to suggest the probable factors of transfer involved at the various stages of mastery. When transfer was made from different stages of partial learning, it appears that general factors may have been largely responsible for the transfer effect, for, after an initial drop, the greater the amount of partial learning the greater the transfer. This trend continued only to a certain point, where apparently specific knowledge elements began to produce interfering effects, and, as a result, the net transfer was reduced. This combination of factors seems to start

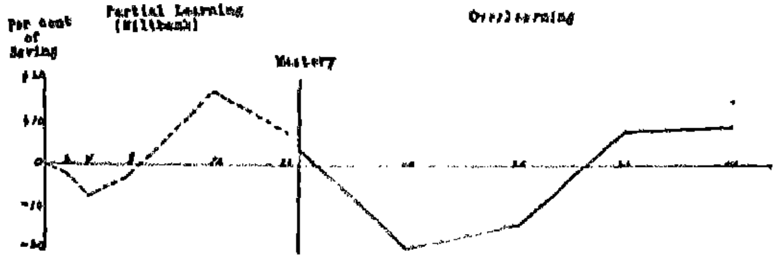

FIGURE 8

SHOWING PERCENTAGE OF SAVING IN TRANSFER FROM DIFFERENT DEGREES OF PARTIAL LEARNING AND OF OVERLEARNING

during the later stage of partial learning and to continue into overlearning. However, specific knowledge elements probably produce greater interfering effects during the early stages of overlearning than at later stages, which would explain the change from negative to positive transfer in the last two stages of overlearning.

CONCLUSIONS

1. The transfer effect was not found to increase consistently with greater amounts of overlearning of the maze when transfer was made from one maze to another. The transfer effect was small and positive when the shift to the second maze was made at mastery. The effect was large and negative when there had been slight overlearning of the first maze, but became positive, although small, when overlearning had been carried further.

2. The specific transfer value from one maze to another was found to differ according to the degree of overlearning of the first maze, as shown by a difference in the distribution of errors. The first habit tended to persist most when overlearning was slight (10-20 trials), it being greater at this stage of mastery than when the shift was made at either mastery or with greater overlearning.

IV

DISCUSSION AND SUMMARY

One point of general interest that might be mentioned is the reliability of the maze as a measure of group differences. Hunter (10) has criticized the maze because he found it to be very unreliable. He concluded that even group differences were not valid. Carr (2), on the other hand, has defended the use of the maze on the ground that inaccuracies of measurement would be positive and negative, and that, in a group, they would tend to cancel one another, leaving the average more or less undisturbed. Recently, Tryon (23) has demonstrated statistically that the maze is not only reliable for group differences but also probably fairly accurate for individual differences. In this connection he developed a more accurate method of comparing group averages. It is well recognized that errors of measurement increase the variability of a set of scores, but errors of measurement occur only when the reliability of the measuring device is not unity. It has been shown by Yule (33), Kelley (12), and others, that the true variability may be found by multiplying the observed variability by the square root of the reliability coefficient ($S.D._t = S.D._{obs.}\sqrt{r}$). If the difference between averages are to be interpreted accurately, the true variability of the various groups should be used. Therefore, true standard deviations should be substituted for observed standard deviations in the reliability of the difference formula, as has been

shown by Tryon (20, 21). When these substitutions are made, the formula is as follows:

$$\text{Critical ratio} = \frac{\text{Difference}}{\sqrt{S.D.^2_{av_1} r_{11} + S.D.^2_{av_2} r_{22}}}$$

The use of the corrected formula operates to increase reliability of the difference. This is as it should be since in the uncorrected formula the measures of variability are too large due to errors of measurement derived from an unreliable test. Tryon (21) has definitely shown that this does not involve an overcorrection, but in populations of the size here involved it is, on the average, slightly too small.

The various methods of computing reliabilities of maze scores have been fully discussed by Tolman and Nyswander (19). The prevalent method, and the one used in the present study, consists in correlating total error scores on odd trials with total error scores on even trials. The correlation coefficient found by this method is the reliability for half the test, since only one-half of the scores were correlated with the other half. But the reliability of the whole maze test is wanted, therefore the Brown-Spearman formula must be applied. These various coefficients, one for each group in each experiment, are shown in Tables 18 and 19. It may be noted that the probable errors of these r's are very large. It was decided, therefore, to use a general reliability coefficient for each of the two mazes. In order to get a representative coefficient for the groups on one maze, an average of the r's for the various groups was computed. But, since a correlation coeffi-

TABLE 18

COEFFICIENTS OF RELIABILITY OF THE MAZE FOR THE VARIOUS GROUPS IN EXPERIMENT I (MAZE X)

Group	Number of animals	Raw r	Corrected r Brown prophecy formula	Probable error of r	z-value for r
1	20	.64	.78	.06	1.06
2	20	.35	.51	.11	.56
3	21	.54	.70	.07	.87
4	19	.55	.71	.08	.89
5	20	.77	.86	.04	1.32
6	16	.70	.83	.05	1.18
7	19	.40	.57	.10	.65
8	18	.88	.94	.02	1.70
9	18	.49	.66	.09	.79

NOTE: Average of the nine z-values is 1.002, which corresponds to an r of .762±.02.

TABLE 19

COEFFICIENTS OF RELIABILITY OF THE MAZE FOR THE VARIOUS GROUPS IN EXPERIMENT II (MAZE Y)

Group	Number of animals	Raw r	Corrected r Brown prophecy formula	Probable error of r	z-value for r
A	20	.89	.94	.02	1.76
B	42	.92	.96	.01	1.91
C & D	24	.83	.91	.04	1.50
E	24	.91	.95	.01	1.84
F	28	.95	.98	.01	2.18

NOTE: Average of the five z-values is 1.84, which corresponds to an r of .951±.006.

cient is not a linear function and is not distributed normally, an ordinary average is not accurate, except when the coefficients are very low. This error may be avoided by use of a method discussed in Fisher (6). This involves the conversion of each r-value into a z-value. The ordinary average of the z's may then be computed and then changed back into an r-value. The average correlation coefficient found in this indirect way is more representative of the group of r's than an

ordinary average would be. The z-values for the various r's are also shown in Tables 18 and 19. The formula for z is as follows:

$$z = \tfrac{1}{2} \log (1+r) - \log (1-r).$$

When this method is applied in the present study, the resulting reliability coefficients are 0.76 and 0.95 for Mazes X and Y, respectively. The fact that the reliability is higher for the more difficult maze is in agreement with the findings of Tolman and Nyswander and Tryon. That Maze Y is more difficult than Maze X is indicated by a *D/SDD* of 5.62 between two Groups, 3 and A, which learned the two mazes under the same conditions. Evidently the mazes used in the present experiment are quite reliable, certainly sufficiently reliable to measure group differences. In general, it appears from the size of the reliability coefficients that the maze compares favorably with many other types of tests as regards accuracy of measurement.

The identical-elements theory of transfer seems to be the prevailing one in recent psychology. However, it should be clear that the results of the present investigation are against this theory as it is generally understood. The usual view seems to be that all transfer effects are dependent upon the presence of identical elements in the two situations. The present study shows that this is not the case, at least in the animal field. In fact, it is evident that such general factors as are involved in the mere handling of animals in which no apparatus is involved are effective in producing transfer in a later situation. The same is true of laboratory

adjustment, and of adjustment to an apparatus which is in no way similar to the test situation. The question may then arise as to whether handling and other such factors are not identical elements, since they occur in both the training and the test situations, and in both cases these factors remain approximately the same. It might be argued that these general conditions are themselves elements. This classification would involve two characteristics—identity and elementary status—which may be regarded as independent of each other. We may admit, without argument, the fact of identity or a high degree of similarity, in most cases at least, between factors in the test and transfer situations. However, we feel that it is important to insist that these factors, in many cases at least, are general, as that term is usually used in psychology. The theory of identical elements seems to consider a response an element no matter what its degree of generality may be.

The responses which are made to handling and the like are more in the nature of an attitude or a general motor set than a specific overt patterned reaction. These responses, obviously, are quite general and constitute what might be considered the background of more specific overt responses. However, the problem of distinguishing this aspect of a response from the specific patterned response is not always an easy one. As a matter of fact, it appears that there are all degrees of responses, since they might be arranged in a continuum from the most general attitude to the most specific overt response. If attitudes of this sort are to be classed as elements, then the identical-elements theory loses its distinctiveness entirely.

Summary

1. A considerable amount of positive transfer was found to occur from situations involving adjustment, handling, and general training to the maze.

2. The amount of transfer was least in the case of general adjustment, somewhat greater under the conditions of handling, and as great from general training as from previous maze training.

3. The locus of transfer induced by general factors is largely limited to the first five or six trials of the second problem, although much less transfer was apparent on the first trial of the test than occurs in maze-to-maze transfer.

4. The transfer effect from maze to maze was sometimes positive and sometimes negative from the different degrees of overlearning, and showed no consistent trend. When transfer was made at mastery, the effect was positive; with transfer at slight overlearning the effect was negative; but with considerable overlearning the transfer effect was positive again.

5. The amount of transfer from maze to maze without regard to direction did not consistently increase with greater amounts of overlearning. The effect was small and positive when transfer was made at mastery of the first maze (4 out of 5 perfect trials), but large and negative with slight overlearning (10-20 trials), and small and positive again with considerable overlearning (30-40 trials).

6. The type of response involved in the transfer effects varied with different degrees of overlearning, as shown by a marked difference in the distribution of

errors in the second maze. The pattern of response established in the first maze tended to persist more when overlearning was slight (10-20 trials) than when it was either greater (30-40 trials) or less (simple mastery, 4 out of 5 perfect trials).

References

1. Bogardus, E. S., & Henke, F. T. Experiments on tactual sensation in the white rat. *J. Anim. Behav.*, 1911, 1, 125-137.
2. Carr, H. The reliability of the maze experiment. *J. Comp. Psychol.*, 1926, 6, 85-93.
3. Cole, L. W. Concerning the intelligence of raccoons. *J. Comp. Neur. & Psychol.*, 1907, 17, 211-261.
4. Dashiell, J. F. Some transfer factors in maze learning in the white rat. *Psychobiol.*, 1921, 2, 329-350.
5. ————. Direction orientation in maze running by the white rat. *Comp. Psychol. Monog.*, 1930, 7, No. 2. Pp. 72.
6. Fisher, R. A. Statistical methods for research workers. Edinburgh: Oliver & Boyd, 1930. Pp. 269.
7. Ho, G. H. Transfer and degree of integration. *J. Comp. Psychol.*, 1928, 32, 397-416.
8. Hunter, W. S. Some labyrinth habits of the domestic pigeon. *J. Anim. Behav.*, 1911, 1, 278-304.
9. ————. Habit interference in the white rat and in human subjects. *J. Comp. Psychol.*, 1922, 2, 29-59.
10. Hunter, W. S., & Randolph, V. Further studies on the reliability of the maze in rats and humans. *J. Comp. Psychol.*, 1924, 4, 431-442.
11. Kinnaman, A. J. Mental life of two *Macacus rhesus* monkeys in captivity. *Amer. J. Psychol.*, 1902, 13, 98-148, 173-218.
12. Kelley, T. L. Statistical method. New York: Macmillan, 1923. Pp. xi+390.
13. McCall, W. A. How to experiment in education. New York: Macmillan, 1923. Pp. 281.
14. Orata, P. T. The theory of identical elements. Columbus: Ohio State Univ. Press, 1928. Pp. 204.
15. Pearson, K. Tables for statisticians and biometricians. Cambridge: Cambridge Univ. Press, 1914. Pp. lxxxiii+143.
16. Razran, H. S. Theory of conditioning and of related phenomena. *Psychol. Rev.*, 1930, 27, 25-43.
17. Richardson, F. A study of sensory control in the rat. *Psychol. Monog.*, 1910, 12, No. 48. Pp. 124.

18. THORNDIKE, E. L. Animal intelligence. New York: Macmillan, 1911. Pp. viii+297.

19. TOLMAN, E. C., & NYSWANDER, D. B. The reliability and validity of maze measures for rats. *J. Comp. Psychol.*, 1927, 7, 425-460.

20. TRYON, R. C. Effect of the unreliability of measurement on the difference between groups. *J. Comp. Psychol.*, 1926, 6, 449-453.

21. ————. Demonstration of the effect of unreliability of measurement on a difference between groups. *J. Comp. Psychol.*, 1928, 8, 1-22.

22. ————. Errors of sampling and of measurement as affecting difference between means. *J. Comp. Psychol.*, 1929, 9, 191-195.

23. ————. Studies in individual differences in maze ability: I. The measurement of the reliability of individual differences. *J. Comp. Psychol.*, 1920, 11, 145-170.

24. WARNER, L. H., & WARDEN, C. J. The development of a standardized animal maze. *Arch. Psychol.*, 1927, No. 93. Pp. 46.

25. WATSON, J. B. Animal education. Chicago: Univ. Chicago Press, 1903. Pp. 122.

26. WEBB, L. W. Transfer of training and retroaction. *Psychol. Monog.*, 1917, 24, No. 104. Pp. 90.

27. WILSON, E. B. Statistical comparison of psychiatric diagnosis in some Massachusetts State Hospitals during 1925 and 1926. *Quar. Bull. Mass. Dept. Ment. Dis.*, 11.

28. WILTBANK, R. T. Transfer of training in the white rat upon various series of mazes. *Behav. Monog.*, 1919-1922, 4, No. 1. Pp. 65.

29. WYLIE, H. H. An experimental study of transfer of response in the white rat. *Behav. Monog.*, 1919, 3, No. 5. Pp. 65.

30. YERKES, R. M. The dancing mouse. New York: Macmillan, 1907. Pp. xxi+290.

31. YERKES, R. M., & COBURN, C. A. A study of the behavior of the pig, *Sus scrofa*, by the multiple-choice method. *J. Anim. Behav.*, 1915, 5, 185-225.

32. YOAKUM, C. S. Some experiments upon the behavior of squirrels. *J. Comp. Neur. & Psychol.*, 1909, 19, 541-568.

33. YULE, G. U. An introduction to the theory of statistics. (8th ed.) London: Griffin, 1927. Pp. xv+422.

DES FACTEURS GÉNÉRAUX DANS LE TRANSFERT DE L'ENTRAINEMENT CHEZ LE RAT BLANC

(Résumé)

Cette enquête a compris deux expériences, dans la première desquelles on a essayé de déterminer la valeur du transfert de l'entraînement général à l'apprentissage du labyrinthe. On a employé des rats blancs mâles comme sujets dans cette expérience, et le labyrinthe a été l'appareil de test employé toujours. On a donné différentes sortes de l'entraînement général sans employer le labyrinthe à neuf groupes de rats, chacun composé d'environ vingt rats, et ensuite on les a entraînés dans un labyrinthe identique. Les situations de transfert pour les différents groupes ont compris l'adjustement général au laboratoire, la manipulation, l'entraînement dans un appareil autre qu'un labyrinthe, et dans un labyrinthe d'une forme différente de celle du labyrinthe du test. Un groupe de contrôle n'avait pas eu d'expérience antérieure, générale ou spécifique, avant ce test du labyrinthe. Les résultats montrent un transfert positif considérable de toutes les diverses situations. L'effet du transfert a été plus grand avec une plus grande quantité de l'entraînement général, l'effet étant aussi grand de l'entraînement dans une boîte à problèmes que de l'entraînement antérieur dans le labyrinthe. On a interprété ces résultats comme de l'évidence que le transfert de l'entraînement peut avoir lieu par des éléments généraux de non-connaissance aussi bien que par des éléments spécifiques de connaissance, et au même degré. Les bons effets de l'entraînement général antérieur ont semblé en grande partie limités aux cinq ou six premières épreuves dans l'apprentissage du labyrinthe. Dans la seconde expérience, on a donné à cinq groupes de rats blancs de différentes quantités de surapprentissage dans le labyrinthe X et ensuite on les a testés dans le labyrinthe Y. On a transféré un groupe au labyrinthe Y après l'arrivée à une simple norme (quatre parcours parfaits sur cinq) dans le labyrinthe X. Les quatre autres groupes ont été transférés au labyrinthe Y après avoir eu dix, vingt, trente, et quarante épreuves de surapprentissage respectivement dans le labyrinthe X. Le groupe de contrôle, bien entendu, a appris le labyrinthe Y sans avoir eu aucune sorte de l'entraînement antérieur dans les expériences. On a trouvé que les effets de transfert des quantités augmentantes de surapprentissage ne montre aucune tendance constante. L'effet net de transfert a été petit et positif pour nul surapprentissage, un peu plus grand et négatif pour une petite quantité (10 à 20 épreuves) de surapprentissage, et moyen et positif pour une plus grande quantité (30 à 40 épreuves) de surapprentissage. On offre des suggestions tentatives pour expliquer ces résultats.

JACKSON

ALLGEMEINE EINFLÜSSE BEI DER ÜBERTRAGUNG DER EINÜBUNG BEI WEISSEN RATTEN

(Referat)

Die vorliegende Untersuchung enthielt zwei Experimente. Das erste war ein Versuch, zu bestimmen, bis zu welchem Grande sich die allgemeine Einübung (general training) auf das Erlernen eines Labyrinthes (maze) übertragen lässt. Man gebrauchte als Versuchstiere bei diesem Versuche *männliche weisse Ratten und als Versuchsapparat diente durchweg das* Labyrinth. Es wurden neun Gruppen von Ratten, jede Gruppe aus 20 Ratten bestehend, auf verschiedene allgemeine Weisen ohne Labyrinth dressiert, und dann an einem immer gleich bleibenden Labyrinth eingeubt. Die zu übertragenden Situationen bei diesen verschiedenen Gruppen umfassten allgemeine Anpassung an das Laboratorium und an die Handhabung (handling) und Einübung an einem Apparat das nicht eine Labyrinth war, *und auch an einem* Labyrinth dessen Gestaltung anders war, als die des eigentlichen Versuchslabyrinthes. Es gab auch eine Kontrollgruppe die keine vorhergehende Einübung genossen hatte, weder allgemeiner noch spezifischer Art, ehe sie an dem Labyrinth gepruft worden war. Die Befunde erweisen eine ziemlich starke positive Uebertragung aus allen verschiedenen Situationen. Desto mehr allgemeine Einubung die Tiere mitgemacht hatten, desto mehr Uebertragung fand statt, wobei die Wirkung der Einübung an einem Aufgabekasten (problem box) ebenso stark war als die einer vorhergehenden Einübung an einem Labyrinth. Der Verfasser deutet diese Befunde als Beweis dafur, dass eine Uebertragung der Einübung eben so gut durch allgemeine, nicht kenntnissartige (non-*knowledge) Einflüsse wie durch spezifische Kenntnisse (specific knowledge* elements) verursacht werden kann. Die vorteilhaften Wirkungen der vorhergehenden allgemeinen Einübung scheinen sich grossenteils auf die ersten fünf oder sechs Versuche bei dem Erlernen des Labyrinthes zu beschränken. In dem zweiten Experiment liess man 5 Gruppen weisser Ratten das Labyrinth X mehr oder weniger übergrundlich erlernen (overlearn). Diese Gruppen wurden dann an dem Labyrinth Y geprüft. Eine Gruppe wurde an das Labyrinth Y versetzt nachdem sie an dem Labyryinth X eine einfache Norm [der Fertigkeit] [4 fehlerfreie Rennungen (runs) aus 5 Versuchen] erzielt hatten. Die 4 anderen Gruppen wurden an das Labyrinth Y versetzt nachdem sie an dem Labyrinth X respektiv 10, 20, 30, und *40 Ueber-erlernungsversuche (overlearning trials) gemacht hatten.* Die Kontrollgruppe erlernte naturlich das Labyrinth ohne vorher irgend eine Art Einübungserfahrung (previous experience training) genossen zu haben. Die Wirkungen der Uebertragung aus steigenden Graden des Uebererlernens zeigten keine beständige Richtung (consistent trend). Die reine (net) Wirkung der Uebertragung war, ohne Uebererlernen, klein und positiv, bei geringem Grade des Uebrerlernens (10 bis 20 Versuche) etwas starker und negativ, und bei höherem Grade des Uebererlernens (30 bis 40 Versuche) mittelmassig stark und positiv. Es werden einige vorlaufige Vorschläge zur erklärung dieser Befunde dargeboten.

JACKSON

$7.00 per volume
Single numbers $2.00

MONTHLY
Two volumes per year

February, 1932
Volume XI, No. 2

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

EDITED BY
CARL MURCHISON

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the post-office at Worcester, Mass., under Act of March 3, 1879

$7.00 per volume
Single numbers $2.00

MONTHLY
Two volumes per year

February, 1932
Volume XI, No. 2

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

THE EFFECT OF COLOR ON VISUAL APPREHENSION AND PERCEPTION*

From the Psychological Laboratories of the University of Minnesota

By

MILES A. TINKER

Accepted for publication by Carl Murchison of the Editorial Board and received in the Editorial Office July 10, 1931.

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the post-office at Worcester, Mass., under Act of March 3, 1879.

I

PROBLEM AND METHOD[1]

There has been an increasing use of color and color combinations in situations designed to convey messages by means of visual symbols. Familiar examples are magazine advertisements, posters, billboards, and automobile license plates. Two important justifications for the use of colors in advertisements and the like are attention value and the pleasant feeling-tone which they arouse. In employing colors for these purposes there is danger that the words or other symbols may lack adequate visibility. Whenever colored letters, words, or other symbols are used on either a white or a colored background, those combinations which, in addition to having attention or affective value, favor quick and accurate apprehension should be chosen. No experiment has yet shown whether attention value or affective value of the colors in which symbols are printed facilitate perception, or investigated certain other influences of color on the apprehension of meaningful characters. Further experimentation and analysis is needed to furnish a more adequate knowledge of the effect of color on visual apprehension and perception.

When background is of a constant color or brightness, colored symbols may be printed in two ways: (1) with all symbols in the same color (homogeneous color series), or (2) with variation in color from one

[1]The expenses of this study were met by a research grant from the Graduate School, University of Minnesota.

symbol to another (heterogeneous color series). The purposes of the present experiment are (1) to determine the effect of color upon visual apprehension of (*a*) homogeneous colored stimuli and (*b*) heterogeneous colored stimuli; and (2) to make a comprehensive analysis of the influence of color and brightness contrast on visual apprehension and perception in reading. The influence of affective value, attention value, and luminosity of colors on the apprehension of colored letters will be considered in the analysis. Other features of the study include determinations of (1) the effects of scoring methods on average span of visual apprehension and on reliability of scores, and (2) the influence of letter position on visual apprehension.

Various ways of measuring span of visual apprehension have been employed. In general, they may be classified into two techniques: (1) studies in which some form of tachistoscope is used to expose the stimuli for an interval which is shorter than the reaction-time of the eye,[2] and (2) experiments in which stimuli are exposed from three to six seconds with or without the aid of some simple apparatus. The latter technique was used in this experiment.

There are two main parts to the investigation. In Part I the influence of homogeneous colors, and in Part II the influence of heterogeneous colors on the apprehension of letters, was studied. With the homogeneous colors 100 sophomore university students served as subjects. There were 50 men and 50 women.

In Part I, the stimulus material consisted of series of

[2]For methods and results see Tinker (26, 28).

colored letters pasted on white cardboard, 7 by 28 inches in size. These letters (ordinary block letters) were 3 inches wide and 4¼ inches high, and were cut from the standard Milton Bradley colored papers. Eight different colors were used: black,[3] orange, violet, blue, red, neutral gray, green, and yellow. Each stimulus card had eight letters, all of the same color. Only consonants were employed and the position of any letter was systematically rotated so that each appeared an equal number of times in all eight positions on the cards. In all, there were 32 stimulus cards, 4 letter series of each color. The colors on successive stimulus cards were systematically rotated throughout the series.

The subjects were tested in groups of about 30 each. The stimulus cards were exposed, one at a time, for a period of three seconds. Exposure was achieved by merely uncovering and covering the stimulus card in a position clearly visible from all parts of the room. Timing was done with a stop-clock. A short practice series preceded the experiment proper.

The subjects were instructed to view carefully the series of letters on the card while it was exposed. As soon as the stimuli were covered the students were to write down as many of the letters as they could remember and in the same order as on the exposure card. Omission of a letter was to be designated by a dash in the series. Motivation appeared adequate and produced excellent cooperation.

In Part II, with the heterogeneous colors, there were

[3]For convenience, all letters are termed colored although two achromatic stimuli (black and gray) are included in the series.

two groups of subjects: 100 men in one and 100 women in the other. All were university sophomores. Scores for both color preferences and visual apprehension were obtained from these subjects.

The same colors were used as in Part I and the stimulus series were constructed in a similar manner, but with the following exceptions: Each letter on any card had a different color, i.e., all eight colors appeared on each card. The colors were systematically rotated so that each color appeared the same number of times (3) in each position on the card. In all there were 24 stimulus series. Presentation of stimuli was the same as in Part I.

Stimuli for color preferences in Part II were made by pasting sheets of colored paper, 5 by 8 inches in size, on white cardboard, 8½ by 11 inches. Each card carried a large identifying number below the sheet of colored paper.

Color preferences were obtained by the method of paired comparisons. Possible space errors were controlled by systematic variation in position of stimuli during succeeding comparisons. The pairs of stimuli were exposed side by side for a three-second interval. From the results an order of preference was obtained for each subject and ranks 1 to 8 assigned the colors. Rank 1 designated the most preferred color.

II

METHODS OF SCORING, RELIABILITY, AND INFLUENCE OF LETTER POSITION

Methods of Scoring

The experimental determinations of the span of visual apprehension (often incorrectly called the span of attention) grew out of the early work of Baxt, Cattell, Wundt, Erdmann and Dodge, and others.[4] Since 1900 there has been an ever-increasing number of studies in this field. The numerous investigators of visual apprehension have not only used various procedures and kinds of stimuli but have also employed a variety of methods in evaluating their data. A survey[5] shows that these procedures differ so greatly that a comparison of the results obtained by different experimenters is hazardous. Three methods of scoring which have frequently been employed are: (1) average spans of visual apprehension computed from raw scores (average number of items correctly reproduced in the proper order per exposure); (2) spans calculated from weighted scores in which full credit is given for correct items and partial credit for partly correct items, i.e., correct symbol but placed in a wrong position; and (3) computing the statistical limen (the stimulus value for which there are correct judgments in 50% of the cases). The last method is considered most adequate by both Dallenbach and Fernberger.[6]

Since no experimenter has used two or more methods

[4]See Tinker (26).
[5]See Guilford and Dallenbach (11).
[6]See Tinker (26).

of scoring the same data, the effect of scoring methods on span of visual apprehension is unknown. Although few would deny that variation in method of scoring alters the span, an experimental determination is needed to yield a quantitative measure of the modification.

In this section a comparison of the spans obtained by scoring the same data by three scoring methods will be made. The three procedures are: (1) average span computed from scores in which a credit of one is given for each item reproduced correctly and in the right place, and, in addition, a credit of one-half for each item reproduced correctly but out of place in the series (Method I); (2) average span calculated from scores in which each item correctly reproduced receives a credit of one irrespective of whether it is or is not in the right place (Method II); and (3) average span derived from scores in which a credit of one is given for each item correctly reproduced and in the right place, but no credit for items reproduced correctly but in wrong positions (Method III).

The analysis of spans derived from these three methods of scoring was made for results from three groups of 100 subjects each. With the homogeneous colored stimuli the group of 100 subjects contained both men and women; with the heterogeneous colored letters one group had 100 men, the other, 100 women.

The average span of visual apprehension was first computed for each individual and then an average of averages for each group as a whole. Comparisons of the latter will be made here.

The means with the standard error of the means as computed by each method and for each group follow:

Group 1

Homogeneous colors, 100 M+W, Method I	Mean = 5.75±.08
Homogeneous colors, 100 M+W, Method II	Mean = 6.22±.07
Homogeneous colors, 100 M+W, Method III	Mean = 5.27±.08

Group 2

Heterogeneous colors, 100 M, Method I	Mean = 5.56±.09
Heterogeneous colors, 100 M, Method II	Mean = 6.08±.08
Heterogeneous colors, 100 M, Method III	Mean = 4.97±.09

Group 3

Heterogeneous colors, 100 W, Method I	Mean = 5.45±.09
Heterogeneous colors, 100 W, Method II	Mean = 5.98±.08
Heterogeneous colors, 100 W, Method III	Mean = 4.98±.10

The first line of the above results should be read as follows: Where homogeneous colored letters were employed as stimuli, and for a group of 100 subjects (men plus women), scoring by Method I (each correct item in proper place = 1, and correct but out of place = ½) yielded an average of 5.75 letters read per exposure; and the standard error of the mean is 0.08. The rest of the results should be read in a similar manner.

The trend of the results are the same for all three groups of subjects. As one might expect, Method II yields a larger average span than I, and Method III has the smallest span of all. It is readily seen that there is a marked tendency to perceive and remember more letters than can be placed in correct serial order. When one-half credit is given for correctly reproduced items that are in wrong places, the average span is increased by approximately one-half an item (Method I *vs.* Method III). Consequently, when full credit is

given irrespective of the order of the reproduced letters, the span is increased by approximately one item on the average (Method II *vs.* Method III). The variability is relatively small in all scoring procedures, but the comparative variability is greatest in Method III for all three groups.

The actual sizes of the differences between averages and the significance of obtained differences are listed below for each of the three groups:

Group 1. Homogeneous Colors (Men and Women)

Method II minus I $= 0.47, \frac{D}{\sigma_D} = 52.57$

Method I minus III $= 0.48, \frac{D}{\sigma_D} = 33.95$

Method II minus III $= 0.95, \frac{D}{\sigma_D} = 23.04$

Group 2. Heterogeneous Colors (Men)

Method II minus I $= 0.53, \frac{D}{\sigma_D} = 14.72$

Method I minus III $= 0.58, \frac{D}{\sigma_D} = 17.58$

Method II minus III $= 1.11, \frac{D}{\sigma_D} = 21.80$

Group 3. Heterogeneous Colors (Women)

Method II minus I $= 0.53, \frac{D}{\sigma_D} = 14.57$

Method I minus III $= 0.47, \frac{D}{\sigma_D} = 21.36$

Method II minus III $= 1.00, \frac{}{\sigma_D} = 23.81$

These results are remarkably consistent from group to group, and the trends hold irrespective of either the color of the stimuli letters, or the sex of the subjects,

The directions of the differences are very stable as shown by the size of the ratio of the difference to the standard error of the difference. The formula for correlated measures was employed to compute the standard error of the difference.

The intercorrelations of the different methods for each of the groups are shown below:

Group 1

Homogeneous colors,	I with II		$r = .994$
Homogeneous colors,	I with III		$r = .984$
Homogeneous colors,	II with III		$r = .869$

Group 2

Heterogeneous colors (men),	I with II	$r = .920$
Heterogeneous colors (men),	I with III	$r = .929$
Heterogeneous colors (men),	II with III	$r = .843$

Group 3

Heterogeneous colors (women),	I with II	$r = .920$
Heterogeneous colors (women),	I with III	$r = .973$
Heterogeneous colors (women),	II with III	$r = .916$

The above results indicate, in general, that Method I correlates very highly with both II and III. In Group 1 these correlations are .994 and .984, respectively; in Group 2, .920 and .929; and in Group 3, .920 and .973. Although the correlations between Methods II and III are lowest in every group (.869, .843, and .916, respectively), they are still rather high. These coefficients of correlation indicate that the span of visual apprehension is being measured with approximately the same degree of adequacy by all three methods of scoring. Method I, however (in which full credit is given for letters reproduced correctly and in their proper place, and one-half credit for items correctly

reproduced but out of place), appears to be slightly more adequate than the others.[7] This is substantiated by the fact that Method I correlates very high with either II or III; it gives some credit for every item that is reproduced correctly; and it penalizes lack of ability to place any letter in its proper place in the series. One is justified, however, in employing either Methods I, II, or III in order to obtain an apprehension score which appears to be more significant for the problem under investigation.

In general, therefore, when it is desirable to know absolute span of apprehension, scoring method is important; but, if size of span in relation to others in the group is wanted, any one of the three methods of scoring may be employed, with a slight preference for Method I.

Reliability of Scoring

The usefulness of any measuring device depends to a considerable degree upon its reliability. In this experiment the method of computing reliability was limited to correlating half *vs.* half or odd *vs.* even scores, since all data from any subject were collected at one sitting. Although a short practice series was given, there was definite adjustment to the experimental situation during the early trials. It was decided, therefore, to compute reliability by correlating the sums of the odd *vs.* the sums of the even scores. The Brown-Spearman "prophecy" formula was then applied to obtain the reliability of the complete test. The raw and

[7]The writer has in progress an experiment to compare the statistical limen with each of the methods of scoring used in the present investigation.

the raised reliability coefficients follow for each of the subgroups (Group 4 is a combination of Groups 2 and 3):

Group 1

Homogeneous colors, Method I $r=.786$, $r_x=.880$
Homogeneous colors, Method II $r=.798$, $r_x=.887$
Homogeneous colors, Method III $r=.768$, $r_x=.869$

Group 2

Heterogeneous colors (men), Method I $r=.868$, $r_x=.929$
Heterogeneous colors (men), Method II $r=.866$, $r_x=.928$
Heterogeneous colors (men), Method III $r=.819$, $r_x=.900$

Group 3

Heterogeneous colors (women), Method I $r=.908$, $r_x=.952$
Heterogeneous colors (women), Method II $r=.850$, $r_x=.919$
Heterogeneous colors (women), Method III $r=.801$, $r_x=.890$

Group 4

Heterogeneous colors (men & women), Method I $r=.860$, $r_x=.925$
Heterogeneous colors (men & women), Method II $r=.855$, $r_x=.922$
Heterogeneous colors (men & women), Method III $r=.805$, $r_x=.892$

The reliability coefficients for Group I are definitely lower than in the other groups (5 to 9 points). The cause of this is probably to be found in the nature of the stimuli in Group 1. All letters on any stimulus card were of the same color. Furthermore, colors on successive cards followed each other in an order which distributed all stimulus cards of any color either in the odd or the even trials. Whatever effect the colors had on apprehension, therefore, influences the sums of the odd more than the sums of the even scores, or vice versa. This effect would tend to lower the odd-even correlation.

Examination of the coefficients in the various subgroups shows that Method I and Method II (in both of which some credit is given for correct reproductions that are in wrong positions in the series) have approxi-

mately the same reliability in each of the groups, although there is some variation from group to group. Another characteristic which is universal in every group is the tendency for Method III to be less reliable by about 6 points than either Methods I or II.

The relative standing in reliability of the three methods is made clearer by examining the scoring methods and by reference to some of the original data sheets. Both Method I and Method II make use of the same data in a slightly different manner (full credit for correctly placed items plus one-half credit for those out of place; and full credit for all items irrespective of order in which reproduced) and should therefore have approximately the same reliability. Method III, however, ignores all letters not reproduced in correct serial order. Reference to the original data sheets reveals some rather striking variability from trial to trial in this method of scoring. Many instances occur in which the first one of two letters of a series are reproduced in the correct sequence and then 3 or 4 letters are reproduced correctly but in a wrong order. Then in the succeeding series all letters reproduced (5 to 7) were in the correct order. Many repetitions of this sort of thing reduces the odd-even reliability coefficient, of course. Logically, it would seem that Methods I and II are bound to yield a more adequate measure of visual apprehension than Method III since they give a weighted credit for partially correct reproduction.

Both the raw and the raised correlation coefficients are high. The raw coefficients range from .768 to .908, with 9 out of 12 above .800. All of the raised

coefficients are approximately .900, the range extending from .869 to .952. These reliabilities are satisfactory for the group comparisons to be made in this investigation.

The data on reliability plus an analysis of the scoring methods justify the conclusion that Method I is the most adequate of these three methods of measuring visual apprehension, that Method II is nearly as good, and that Method III is least adequate. The relatively high intercorrelations between methods, and the size of the reliability coefficients justify the added conclusion that any one of the scoring methods may be used and still achieve an adequate measure of visual apprehension.

Influence of Letter Position on Visual Apprehension

As every one knows, Western Europeans commonly read symbols (words, numbers, etc.) from left to right in a line of print. By the time one has reached the sixth grade or higher, this habit is rather rigidly fixed. It is only to be expected that adults, in apprehending tachistoscopically exposed series of characters, either in sense or in nonsense arrangements, tend to read from left to right and to reproduce correctly more characters at the left end of the series. The influence of character position on visual apprehension probably varies somewhat, however, with length of series as well as with uniformity of length of series.

Crosland and Johnson (6) report data which indicate that, in series of letters, the position at the left

end is the most favorable position in the series for correct apprehension, and that each succeeding position toward the right is less favorable than the location at its left. The authors state that "the percentage of letters correct in all respects decreases gradually and consistently from left to right." They employed letter series ranging from 3 to 10 items in length. Various studies of visual apprehension show that series 3 to 5 items in length are frequently apprehended entirely correctly (5, 27).

The results of Crosland and Johnson would have been more adequate if they had analyzed each letter series of a given length by itself. Even then, the shorter series would not give an accurate picture of the influence of letter position unless the length of the series represented approximately the average span of the individual. Any test that is too easy does not measure discriminatively.

The writer considers that to obtain an accurate picture of the relative influence of letter position on range of visual apprehension, the letter series should be uniform in length and equal or slightly greater than the subject's average span. If these conditions are fulfilled it is probable that results somewhat different from those of Crosland and Johnson will be found. Although these authors state that there was a "strong tendency in all subjects to succeed in catching the last letter on the card" (at the right), their method of computing the percentage of maximum score correct does not reveal this trend. Also in Crosland's supplementary graph (5) there appears to be too rapid a drop in the curve

for earned absolute scores (and also for earned percentage scores) in the first few letter positions.

Because of the inconclusive evidence available, the responses of 100 men and of 100 women to 24 series of 8 letters each were scored in two different ways and the influence of letter position on visual apprehension determined. This was the series of heterogeneous colored stimuli described in Chapter I. The two methods of scoring used have been described above: (1) In Method II, every letter reproduced correctly was given a credit of 1 irrespective of the order of reproduction. For the present purpose each correctly reproduced letter was credited to the position it held in the stimulus series. (2) In Method III a credit of 1 was given to each letter reproduced correctly and in the right place. No credit was given for misplaced letters.

Since the average span of the group for these series was between 5 and 6 letters, the eight-letter series was slightly greater than the average. For individual members of the group the average spans ranged from approximately 3.5 to 7.5. Inspection of the original data sheets revealed that all eight letters of a series were reproduced correctly only infrequently.

The average number of times that the letters occurring in each position were apprehended correctly was computed by scoring Methods II and III for 100 men, for 100 women, and for the combined group. The highest possible score for any position is 24. The results for the two subgroups are given in Table 1.

Inspection of the first row in the table shows that

TABLE 1

THE INFLUENCE OF LETTER POSITION ON THE RANGE OF VISUAL APPREHENSION

Highest Possible Score = 24

	100 Men				100 Women			
Position in	Method II		Method III		Method II		Method III	
series	Mean	$\sigma_{Dist.}$	Mean	$\sigma_{Dist.}$	Mean	$\sigma_{Dist.}$	Mean	$\sigma_{Dist.}$
1*	23.0	1.5	22.6	2.6	22.8	1.6	22.3	1.9
2	22.4	2.1	20.5	2.7	22.2	1.9	20.2	2.8
3	21.7	2.5	19.0	3.3	20.8	2.4	17.9	3.3
4	20.5	2.7	17.6	3.7	19.8	3.1	16.9	4.3
5	17.2	3.3	12.8	4.2	15.9	3.9	12.1	4.8
6	15.8	4.3	10.4	4.8	15.4	4.4	10.4	5.0
7	11.9	6.3	7.6	5.5	12.4	5.2	8.5	5.2
8	12.5	6.4	9.2	6.1	13.6	5.5	10.3	5.5

*Letter position number 1 is at the left end of the series.

letters in position 1 are apprehended correctly approximately 95% of the time (score of 24 = 100%), irrespective of scoring method or the sex of subjects. Changes in scores with changes in letter position reveal similar trends for men and for women, but the scores for women are slightly smaller than for men except in positions 7 and 8 where they are slightly greater. In general, there is in both methods of scoring a decrease in score from each letter position to the succeeding one through position 7. In position 8 there is always a slight increase in average score.

When results for the men and the women are combined the average scores for succeeding letter positions are:

Letter Position	1	2	3	4	5	6	7	8
Method II	22.9	22.3	21.3	20.2	16.6	15.6	12.1	13.1
Method III	22.5	20.3	18.4	17.2	12.4	10.4	8.1	9.7

There are variations in constancy of trends from one scoring method to the other. These trends are best

shown by noting the decrease or increase in average score from one letter position to another in each method of scoring. Differences obtained by subtracting the score in one position from that in the preceding are given below for the combined group (100 men plus 100 women) as well as for each sub-group (to be read: score in position 1 minus score in position 2 equals 0.6, etc.):

Differences between:	1-2	2-3	3-4	4-5	5-6	6-7	8-7
Men, Method II	0.6	0.7	1.2	3.3	1.4	3.9	+0.6
Women, Method II	0.6	1.4	1.0	3.9	0.5	3.0	+1.2
Men, Method III	2.1	1.5	1.4	4.8	2.4	2.8	+1.6
Women, Method III	2.1	2.3	1.0	4.8	1.7	1.9	+1.8
M + W, Method II	0.6	1.0	1.1	3.6	1.0	3.5	+0.9
M + W, Method III	2.2	1.9	1.2	4.8	2.0	2.3	+1.6

In either method of scoring these data show the same tendency with men subjects, with women subjects, and with men and women combined. This demonstrates a consistency of general trend from group to group. The following discussion is based on the results for the combined group ($M + W$).

While both methods show a decrease in score at each succeeding position from position 1 through 7 and then an increase at 8 (indicated by +), the amount of change is always less in Method II, with the single exception of sixth to seventh position where it is greater in Method II. This difference in amount of change is probably due to method of scoring. It will be remembered that Method II gave credit for reproduced letters irrespective of order of reproduction.

Table 1 shows that scores in II are slightly greater than in III for every letter position. It is evident that ability to reproduce letters in their correct sequence decreases more rapidly from position to position than ability to reproduce letters irrespective of sequence.

In Method II there is a slight decrease in score from position 1 to 2, 2 to 3, and 3 to 4. From 4 to 5 there is a decided decrease, more than 3 times the preceding amounts. Then comes a lesser decrease from 5 to 6, another large decrease from 6 to 7, and, finally, a small *increase* from 7 to 8. In comparison with II, Method III reveals much larger decreases from position 1 to 2, 2 to 3, 4 to 5, and 5 to 6 and a greater increase from 7 to 8. Positions 6 to 7 show, however, a smaller decrease in Method III. Both methods show a marked drop in score from position 4 to 5. The same is true for position 6 and 7. The amount of change in score from positions 1 to 2 and 7 to 8 contrast markedly in the two methods of scoring for, as shown above, the decrease is much less in Method II.

The findings in this study differ somewhat from those reported by Crosland and Johnson (6) and by Crosland (5). The former reported that "the percentage of letters correct in all respects decreases gradually and consistently from left to right." Curves based on their data, but printed in the latter study show a rather rapid drop from position 1 through 5, and then a much slower drop from position 5 through 10. The rapid drop in the first half of these curves (5, p. 377) apparently is a function, in part at least, of the length of the stimulus series. The inclusion of series of 3, 4, and

5 letters with the longer ones, as stated in another discussion (27) undoubtedly favored the accumulation of high frequencies of correct responses in the first 4 or 5 positions in the letter series. Such an effect would increase the rapidity of the drop in the first part (left half) of the curves.

In the present investigation, curves drawn from the data in Table 1 would contrast markedly with those of Crosland. Although the drop in average score in our study is somewhat more rapid in Method III than in II, both show a consistent and very gradual drop from letter position 1 to 4. From position 4 to 5 and 6 to 7 occur marked drops followed by a rise from 7 to 8. While the exposure interval in the present experiment was 3 seconds and in Crosland's 150 σ, the writer considers the differences in results to be largely due to the use of stimulus series of various lengths in the latter study and series of a constant length, which was slightly above the average span of the subjects, in the other investigation.

The change in average score in successive letter positions in our study may be explained as follows: The accuracy of apprehension of the first 3 or 4 letters in any series was high for all subjects. This, combined with the habitual tendency to apprehend any series of characters from left to right, led to a rather gradual decrease in letters correctly apprehended. Since the average for the group lies between 5 and 6 letters, a concentration of individual average spans occurs around 5. This led to an increased concentration of mistakes at position 5. In position 6, which is also

close to the group average, one would expect only a few more mistakes than at 5. Position 7, however, is definitely beyond the span of all but a very small percentage of the group, and consequently another marked drop in average score is discovered. The slight increase of correct reproduction of letters in position 8 is due to a tendency manifested by many of the subjects to apprehend the last item of a series even though they had missed two or three letters preceding it. Evidently the end position favors more accurate reproduction than the adjoining location just within either end of the series.

One is justified in concluding that letter position has a definite effect on visual apprehension. From left to right there is a decrease in the average number of letters correctly reproduced in each succeeding letter position through the seventh and then a slight increase in the last position. The decrease from the first to the fourth is constant and gradual. There are rapid drops in score from the fourth to fifth and sixth to seventh positions, and always an increase at the eighth position. These produce marked irregularities in the consistency of trend from position to position in the series. Both absolute and relative variability of scores increases consistently from letter position 1 through 8 (Table 1).

III

EFFECT OF COLOR ON APPREHENSION OF LETTERS

In studying the effect of color on the apprehension of letters printed on a uniform background, two arrangements of stimuli must be considered: (1) series in which all letters on a single stimulus card are of the same color, i.e., homogeneous colors; and (2) variation of color from one letter to another on the same stimulus card, i.e., heterogeneous colors. There is considerable difference between the two situations. In the former no difference between hue or brightness of color occurs from symbol to symbol in any single series of letters. There is with each succeeding symbol in the latter, however, a simultaneous change in both hue and brightness. Influence of color on apprehension may be different in the two situations.

The object of the investigation reported in this chapter was to determine the effect of both homogeneous and heterogeneous colors on visual apprehension of letters when background of stimuli is constant (white).[8]

In the analysis of results, color preference, attention value, and luminosity, as well as hue of the colors, have been considered as possible factors involved in producing differences in apprehension.

[8]When the color of both background and symbols are varied a somewhat different situation is produced. The effect of color combinations on perception has been investigated in another experiment which will be reported in Chapter IV.

HOMOGENEOUS COLORS

In this part of the study, 32 series of colored stimuli, each containing 8 letters, were read. There were 4 series of each of the following eight colors: violet, red,

TABLE 2
VISUAL APPREHENSION OF HOMOGENEOUSLY COLORED STIMULI
$N = 100$ (50 men + 50 women)

Color	Rank	Mean	σ_M
Violet	1	6.50	0.09
Red	2	6.43	0.09
Green	3.5	6.34	0.10
Gray	3.5	6.34	0.11
Orange	5	6.31	0.10
Black	6	5.97	0.09
Blue	7	5.93	0.08
Yellow	8	5.70	0.09
Average*	——	6.22	0.07

*This average was computed from the original data.

green, gray, orange, black, blue, yellow. One hundred subjects (50 men and 50 women) observed in the experiment. The procedure has been described in Chapter I.

The average score (span) of each subject for each color was computed by Scoring Method II in which each letter correctly reproduced was given a score of one, irrespective of the order of reproduction. Each group mean, therefore, is an average of 400 readings. In scoring for effect of color on apprehension, it is necessary to give equal weight to each reproduced letter, as is done in Scoring Method II. As noted in Chapter II, when all 32 series of stimuli are considered, this method of scoring has a reliability coefficient of .89. When only 4 series (one color) are employed, the consistency of performance is considerably lessened. This reliability is represented by a co-

efficient of .55, which is large enough, however, to justify group comparisons of the data involved.

Table 2 contains the basic data for visual apprehension of homogeneously colored letters. Column 1 shows the various colors used, Column 2, the relative rank of mean span of apprehension listed from largest to smallest, Column 3, the mean number of letters correctly apprehended per exposure, and Column 4, the standard errors of the means. The latter are all comparatively small and show little variation from color to color, the range being from 0.08 to 0.11.

Examination of Column 3 reveals that the mean spans fall into 4 groups. At the top of the list stand violet and red with average spans of 6.50 and 6.43, respectively. In the next lower group, green, gray, and orange have approximately equal spans. The means are 6.34, 6.34, and 6.31, respectively. There is then a considerable gap to black and blue, which stand close together with average spans of 5.97 and 5.93. Finally, yellow stands at the bottom of the list with a span of only 5.70. The last row of the table gives the total average of 6.22 for all colors (32 series per subject).

The ranking of the colors in Column 2 is apt to be somewhat misleading unless the differences between averages are considered. Since green and gray differ from each by only .003, they have been given the same rank, but orange, which differs from either green or gray by only .03 receives a different rank. The ranges of differences between mean spans for each color in comparison with the others are given below:

Violet	.07 to .80	Orange	.03 to .61
Red	.07 to .73	Black	.04 to .53
Green	.003 to .64	Blue	.04 to .57
Gray	.003 to .64	Yellow	.23 to .80

An adequate interpretation of these differences rests on a knowledge of the ratios of the differences to the standard error of the differences. These ratios are listed in Table 3. The formula for correlated measures was employed to compute the standard errors of the differences since the spans for the various colors correlated with each other. These coefficients of correlation ranged from .40 to .66, with an average of .55.

TABLE 3

THE RATIOS $\frac{D}{\sigma_D}$ FOR THE DIFFERENCES BETWEEN THE AVERAGES GIVEN IN TABLE 2

	Violet	Red	Green	Gray	Orange	Black	Blue	Yellow
Violet		0.93	1.94	1.56	2.25	6.85	6.77	9.88
Red	0.93		1.03	0.82	1.34	5.58	5.43	8.23
Green	1.94	1.03		0.03	0.33	4.78	4.77	6.74
Gray	1.56	0.82	0.03		0.23	3.27	4.00	5.94
Orange	2.25	1.34	0.33	0.23		4.45	4.27	7.07
Black	6.85	5.58	4.78	3.27	4.45		0.40	3.17
Blue	6.77	5.43	4.77	4.00	4.27	0.40		2.83
Yellow	9.88	8.23	6.74	5.94	7.07	3.17	2.83	

Seventeen of the 28 differences between means are of relatively high significance. Sixteen have a $\frac{D}{\sigma_D}$ of 3.17 or greater and the seventeenth has a ratio of 2.83. One important trend in Table 3 should be noted. The average span for yellow, which is by far the smallest of the eight, consistently shows a statistically very reliable difference from all the other means. Black and

blue are the only other two means which reveal a large percentage of highly reliable differences. Each of the others (violet, red, green, gray, and orange) have only three differences of relatively high statistical reliability. A difference had to be about .25 or above for the $\frac{D}{\sigma_D}$ to equal 3.00 or more. There are, as demonstrated by these results, significant differences between apprehension scores of certain colored letters.

Among the factors which might have influenced apprehension of the colored letters, color preference, attention value, and luminosity should be listed. If the rankings for preference and attention value obtained with other subjects than those employed in this experiment are accepted as approximately valid, comparisons may be made with our data. Ranking for color preference was obtained from a combined group of 100 men and 100 women (see *Heterogeneous Colors* below), and that for attention value from a group of 119 men and women (Adams, 1, pp. 118-119). These rankings, together with that for luminosity [determined by DeCamp's method (7)], are given in Table 4.

TABLE 4

RANKINGS FOR SPANS OF COLORED LETTERS, COLOR PREFERENCE, ATTENTION VALUE, AND LUMINOSITY OF COLORS

Span		Preference		Attention V		Luminosity	
Violet	(1)	Blue	(1)	Orange	(1)	Yellow	(1)
Red	(2)	Green	(2)	Red	(2)	Green	(2)
Green	(3.5)	Red	(3)	Blue	(3)	Orange	(3)
Gray	(3.5)	Orange	(4)	Black	(4)	Gray	(4)
Orange	(5)	Violet	(5)	Green	(5)	Blue	(5)
Black	(6)	Yellow	(6)	Yellow	(6)	Violet	(6)
Blue	(7)	Black	(7)	Violet	(7)	Red	(7)
Yellow	(8)	Gray	(8)	Gray	(8)	Black	(8)

In the first column of Table 4, reading from top to bottom, are given the rankings for spans of visual apprehension for colored letters. The numbers in parentheses, following the colors, designate the rank. Violet has the largest span and is given a rank of 1. In Column 2, blue is the most preferred color; in Column 3, orange has the greatest attention value; and in Column 4, yellow possesses the greatest luminosity. By luminosity is meant the percentage of white in the color. For example, white has 100 per cent luminosity; black, zero per cent.

A comparison of the relative positions of each color in the successive columns reveals a striking lack of relationship between most of the rankings. There appears to be a trend toward a negative relationship between span for colored letters and luminosity, and toward a positive relation between preference (affective value) and attention.[9] The relationships are more adequately revealed by correlations between the rankings. They are given below:

Color span	*vs.* color preferences	$\rho = +.021$
Color span	*vs.* attention value	$\rho = -.153$
Color span	*vs.* luminosity	$\rho = -.351$
Preferences	*vs.* attention value	$\rho = +.527$
Attention value	*vs.* luminosity	$\rho = -.190$

The first correlation coefficient of .021 shows that no relation exists between size of span for letters of a certain color and preference for that color.

There appears to be a very slight tendency for colors

[9]Only slight sex differences exist in either attention value or affective value of colors. The ranking for men correlates +.762 with that for women for the former and +.976 for the latter (see *Heterogeneous Colors* below).

of high attention value to coincide with colored letters having a short span of visual apprehension. The coefficient of —.153 is so small, however, that it could be more adequately interpreted as signifying no relationship between attention value and span.

The coefficient of —.351 indicates an appreciable tendency for high percentage of luminosity to be accompanied by small span for letters of a given color. This trend is easily discovered in Table 4. Yellow, the most luminous color, has the shortest span. And violet, which has the greatest span, is sixth in luminosity, close to the non-luminous end. Furthermore, red, which has next to the largest span, is next to the least luminous color.

All the colored letters were on a white background. The less the luminosity of a color, therefore, the greater the contrast in brightness between color and background. Yellow, with approximately 76 per cent luminosity, shows small contrast with white, but violet and red, with about 8 and 5 per cent, respectively, produce marked contrast with the background. Apparently this difference of contrast in brightness between color of letters and background affects span of apprehension to a certain degree. Black (with zero per cent luminosity) and blue are the only colors markedly misplaced in the tendency for size of span to correlate negatively with luminosity. If black and blue are excluded from the series, the correlation between size of span and percentage of luminosity becomes —.901. As one would expect, the greatest effects on apprehension occur with colors of very high or very low percentage of

luminosity, i.e., violet and yellow, especially the yellow, where span differs significantly from all the others. Green, although apparently somewhat displaced in a rank correlating negatively with luminosity, has an average which differs from orange by only .03 of a letter. Its actual displacement is therefore slight.

One is justified in concluding that luminosity of colors has some effect on apprehension of homogeneously colored letters. In general, the greater the luminosity the smaller the span of apprehension. Hue of colors alone probably has little effect on visual apprehension score. The size of the spans for blue and black, which represent marked exceptions to this generalization, are unexplainable, either in terms of the method of the experiment or from other data at hand.

Color preference correlated with attention value yielded a coefficient of .527, indicating a fair amount of relationship. Attention value, however, correlates with luminosity by only —.190. These relationships will receive further attention in our discussion of heterogeneous colors.

Heterogeneous Colors

In this part of the investigation, 24 series of colored stimuli, each containing 8 differently colored letters, were read. They were the same colors employed in the homogeneous series. Each color appeared a single time on each stimulus card. Arrangement of colors, letters, and procedure of experimentation have been described in Chapter I.

Span of visual apprehension for heterogeneously

colored stimuli has high reliability by three methods of scoring, as has been shown in Chapter II. To discover the effect of color on visual apprehension with heterogeneously colored stimuli, it was necessary to score the responses somewhat differently than with homogeneous colors. Since each color occurred once on every stimulus card, and an equal number of times (3) in each letter position during the 24 series, the total number of letters reproduced of a given color will yield a measure of apprehension for that color. A comparison of these measures for the various colors will demonstrate any effect which color has on the apprehension of heterogeneously colored letters.

The responses were scored for the number of times letters of each color were reproduced in the course of the 24 series. A letter was called correct, irrespective of sequence of reproduction. The highest possible score was 24, since each color appeared only once in each series.

TABLE 5

APPREHENSION OF HETEROGENEOUSLY COLORED LETTERS BY 100 MEN, BY 100 WOMEN, AND BY THE COMBINED GROUP

Color	*100 Men* Mean	σ_M	*100 Women* Mean	σ_M	*Men+Women* Mean	σ_M
Black	18.67 (1)*	.30	18.14 (3)	.29	18.40 (1)	.21
Orange	18.40 (2)	.27	18.20 (1)	.27	18.30 (2)	.19
Violet	18.34 (3)	.29	17.77 (6)	.30	18.06 (4)	.21
Blue	18.31 (4)	.29	18.16 (2)	.27	18.24 (3)	.20
Red	18.23 (5)	.29	17.82 (5)	.29	18.02 (5)	.21
Gray	17.90 (6)	.28	17.98 (4)	.29	17.94 (6)	.20
Green	17.77 (7)	.28	17.37 (8)	.29	17.57 (7)	.20
Yellow	17.36 (8)	.31	17.53 (7)	.36	17.44 (8)	.24

*The numbers in parentheses indicate the ranks from the highest to the lowest average.

The means and the standard errors of the means for the number of times letters of each color were apprehended by men, by women, and by the combined group are given in Table 5. The numbers in parentheses following each mean indicate the relative rank (1=highest average). Columns 3, 5, 7 of the table show that variability of the means is relatively constant from color to color.

As will be shown later, there are no sex differences in span of visual apprehension. When the responses are scored for colors apprehended, however, the colors change rank somewhat from men to women. For men the rank order from highest to lowest mean score is black, orange, violet, blue, red, gray, green, yellow; while for women it is orange, blue, black, gray, red, violet, yellow, green. The greatest discrepancy between men and women, however, is only 3 ranks (violet). The correlation between the ranks of men and of women is +.714. The rest of the discussion will be based on the results of the combined group shown in Columns 6 and 7 (N=100 men+100 women), since, in general, the results for both men and women reveal the same trends.

The effect of color on apprehension for the combined group is shown by the differences between the means in Column 6 of Table 5. The range of these differences for each color in comparison with all the others is given below:

Black	.10 to .96	Red	.08 to .58
Orange	.06 to .86	Gray	.08 to .50
Violet	.04 to .62	Green	.13 to .83
Blue	.06 to .80	Yellow	.13 to .96

These differences range from .04 to .96. While many

of the differences are small, some are large enough to yield comparatively marked stability for the direction of the difference.

The significance of each difference has been evaluated by computing the ratio of the difference to the standard error of the difference. These ratios (for combined group) are found in Table 6. The colors in Column 1 have been listed according to size of average score with the highest apprehension score (black) at the top. The standard error of the difference was computed by the formula for correlated measures. The ranges of the intercorrelations between the average number of letters apprehended in each color are listed below for men, women, and the combined group:

TABLE 6

THE RATIOS $\frac{D}{\sigma_D}$ FOR THE DIFFERENCES BETWEEN THE AVERAGES GIVEN IN COLUMN 6 OF TABLE 5

	Black	Orange	Blue	Violet	Red	Gray	Green	Yellow
Black		0.60	0.88	1.98	2.17	2.88	4.93	4.30
Orange	0.60		0.38	1.42	1.69	2.18	4.50	4.20
Blue	0.88	0.38		1.03	1.16	2.03	3.81	4.11
Violet	1.98	1.42	1.03		0.23	0.70	2.79	2.97
Red	2.17	1.69	1.16	0.23		0.49	2.80	2.60
Gray	2.88	2.18	2.03	0.70	0.49		2.23	2.36
Green	4.93	4.50	3.81	2.79	2.80	2.23		0.60
Yellow	4.30	4.20	4.11	2.97	2.60	2.36	0.60	

100 men range = .52 to .72, average = .63
100 women range = .42 to .72, average = .60
100 men + women range = .50 to .72, average = .62

These correlations indicate that the relative number of letters apprehended remains fairly consistent from color to color.

The trend of magnitude of differences and reliability

of differences is the same for men and women as for the combined group. The conclusions from results for the combined group, therefore, may be considered to hold for a group of either 100 men or 100 women.

The most striking trend found in Table 6 is the consistently high significance of the differences between average scores of both green and yellow and all the other colors. Chances that the direction of the differences will be reversed on future experiments are very slight indeed. Approximately a chance difference exists between scores for green and yellow. Reference to Column 6 of Table 5 shows that yellow and green have markedly smaller averages than the other colors, and that the difference between green and yellow is very small (0.13).

As with homogeneous colors, to analyze the trends revealed in Tables 5 and 6, one should consider the possible effects of affective value (color preferences), attention value, and luminosity of colors on the apprehension of colored letters. Color preferences were obtained for both men and women along with the visual apprehension scores. The orders of preference are given below for 100 men, 100 women, and the combined group of men plus women:

Men	*Women*	*Men and Women*
Blue	Blue	Blue
Green	Green	Green
Red	Orange	Red
Orange	Red	Orange
Violet	Violet	Violet
Yellow	Yellow	Yellow
Black	Black	Black
Gray	Gray	Gray

Blue was most preferred by both men and women, and by the combined group. There are practically no sex differences in affective value for these colors (Milton Bradley's standard colors). Red and orange are the only colors which occupy non-corresponding ranks in the two lists. The correlation between the ranking for men and women is +.976. This coefficient is also a measure of the stability of the ranking for the combined group. The orders of preference shown above are approximately the same as those given by other investigators. Color preferences are stable and lasting and apparently are not dependent to any marked degree upon the temporary condition of the person making the choice.[10]

For convenience of comparison the eight colors are ranked for average number of letters apprehended, preference, attention value, and relative luminosity in Table 7. Black had the largest average number of let-

TABLE 7

RANKINGS FOR AVERAGE NUMBER OF LETTERS APPREHENDED, COLOR PREFERENCE, ATTENTION VALUE, AND LUMINOSITY OF COLORS

Apprehension average	Color preference	Attention value*	Relative luminosity
Black	Blue	Orange	Yellow
Orange	Green	Red	Green
Blue	Red	Blue	Orange
Violet	Orange	Black	Gray
Red	Violet	Green	Blue
Gray	Yellow	Yellow	Violet
Green	Black	Violet	Red
Yellow	Gray	Gray	Black

*Adams (1, Table 4).

[10]For a summary of results on affective value of colors see Poffenberger (20, pp. 438-447).

ters apprehended, blue was the most preferred color, orange had the highest attention value, and yellow is the most luminous color.

Certain trends are noticeable from mere inspection of the successive rankings. There appears to be no relation between the rankings for apprehension and for preference. Apparently, however, there is some correspondence between ranking for apprehension and attention value, for there are only slight differences between ranks in the two series for orange, blue, green, yellow, and gray. But between ranking for apprehension and luminosity the relationship appears to be an inverse one as indicated by the positions of yellow, green, and black. These relationships are given mathematical expression in the correlations which follow. Coefficients of correlation are given for men and for women as well as for the combined group of men plus women. Similar trends in all subgroups will allow conclusions to be drawn with greater confidence.

Colors apprehended *vs.* color preferences (men)	$\rho = -.071$
Colors apprehended *vs.* color preferences (women)	$\rho = +.095$
Colors apprehended *vs.* color preferences (M+W)	$\rho = +.024$
Colors apprehended *vs.* attention value (men)	$\rho = +.429$
Colors apprehended *vs.* attention value (women)	$\rho = +.190$
Colors apprehended *vs.* attention value (M+W)	$\rho = +.524$
Color preferences *vs.* attention value (men)	$\rho = +.524$
Color preferences *vs.* attention value (women)	$\rho = +.667$
Color preferences *vs.* attention value (M+W)	$\rho = +.527$
Colors apprehended *vs.* luminosity (men)	$\rho = -.690$
Colors apprehended *vs.* luminosity (women)	$\rho = -.337$
Colors apprehended *vs.* luminosity (M+W)	$\rho = -.667$
Attention value *vs.* luminosity (men)	$\rho = -.119$
Attention value *vs.* luminosity (women)	$\rho = -.071$
Attention value *vs.* luminosity (M+W)	$\rho = -.130$

Color preferences *vs.* luminosity (men)	$\rho = +.095$
Color preferences *vs.* luminosity (women)	$\rho = +.105$
Color preferences *vs.* luminosity (M+W)	$\rho = +.095$

The coefficients of —.071, +.095, and +.024 demonstrate a complete absence of relationship between apprehension of colored letters and color preference. Immediate memory (visual apprehension) of colored letters, therefore, is not affected by the pleasant or unpleasant feeling tones which usually accompany the perception of the colors.

It is often stated that one tends to recall the pleasant more readily than the unpleasant. There is, however, considerable disagreement concerning the influence of feeling tone on memory. Tait (24) found that pleasant colors were recognized more frequently than unpleasant or indifferent ones. Although our results do not agree with Tait's, they are supported by the experimental findings of Gordon (9, 10) and Anderson and Bolton (2). In her first study, where the procedure was similar to ours, Gordon attempted to determine whether the pleasantness or unpleasantness accompanying the perception of simple geometric figures and colors has any influence on the accuracy of memory for these experiences. The observers, after viewing more or less complex figures for three seconds, described the pictures immediately after they were removed from vision. Each figure was classed as pleasant, unpleasant, or indifferent. In another part of the experiment, the observers viewed for one second figures consisting of 9 colored squares. At the end of the exposure the observer reported the affective value of the object and

named the colors seen. The immediate memory of both geometric figures and colored objects was practically the same for pleasant, unpleasant, and indifferent figures. In delayed memory (3 weeks) recognition was equally good for figures of all degrees of pleasantness. The conclusion is that there is no direct influence of feeling tone on memory.

In her second study, Gordon discovered no relationship between affective value and immediate memory for odors. The correlation between ranks for affective value and memory was —.07. Anderson and Bolton also employed odors for stimuli and found no significance between immediate memory of pleasant and unpleasant stimuli. Many other experiments have studied the influence of feeling tone on memory for various types of experiences. Certain investigators claim to have demonstrated the presence, others the absence, of such influence.

The present study corresponds most nearly to Gordon's first experiment. The principal difference was the determination in our experiment of a rank order for the affective value of colors appearing on the stimulus cards. Our results correspond to those of Gordon and others who have found no influence of affective value on immediate memory.

The coefficients of +.429, +.190, and +.524 between apprehension score and attention value demonstrate that attention value is potent to some degree in determining apprehension of heterogeneously colored letters. The tendency is for colors of greater attention value to have larger apprehension scores on the aver-

age. Reference to Table 7 reveals that both orange and blue rank high in apprehension score and attention value, while both yellow and gray rank low.

Rankings for color preference correspond somewhat to those for attention value. The correlations are +.524, +.667, and +.527, showing that most preferred colors tend to have greater attention value. Hence, while related to attention value, color preference does not influence apprehension of colored letters, although attention value apparently does to a certain degree.

In general, the greater the luminosity of a color the smaller the apprehension score. This is shown by the correlations between luminosity and apprehension. The coefficients are —.690, —.337, and —.667. This relationship is apparently due to the variation of brightness contrast between colored letters and white background (cardboard). The brightness contrast ranges from 100 per cent for black letters on white background to approximately 24 per cent for yellow on white. The large amount of contrast for black letters on white cardboard results in the apprehension of the largest number of letters. On the other hand, the relatively small contrast between either yellow or green in comparison with the white results in a much smaller apprehension score than with any of the other colors (see Tables 5 and 7).

Although there is less relationship between apprehension scores and both attention value and luminosity for women than for men, the effect of this sex difference is not apparent with the combined group of men

plus women. The larger group yields practically the same correlation as the men.

Since the correlations between attention value and luminosity are so small, there is probably little or no relationship present. The coefficients are —.119, —.071, and —.130. Similarly, color preferences appear unrelated to luminosity. The coefficients are +.095, +.105, and +.095.

The above analysis of the data presented in Tables 5, 6, and 7 leads to the conclusion that at least two agents are operating to produce differences in the average number of letters apprehended for various colors in heterogeneous color series. These factors are the relative luminosity and the attention value of the colors. There is a tendency for less luminous colors to have higher apprehension scores. This is due apparently to the brightness contrast between color and white background. An important exception to this trend is the score for orange. Orange has relatively high luminosity which means small brightness contrast between color and background, and also a high apprehension score. The large apprehension score of orange is to be explained in terms of attention value, since it has greater attention value than any other color and there is an appreciable correlation between attention value and apprehension score for colored stimuli. The affective value of colors has no influence on apprehension of colored stimuli (i.e., on immediate memory for briefly exposed objects).

Homogeneous versus Heterogeneous Colors

The above discussion indicates that color has a somewhat different effect on visual apprehension with heterogeneously than with homogeneously colored stimuli. With homogeneous colors, the effects appear to be inconsistent to a considerable degree. The spans of apprehension for violet, red, green, gray, and orange group themselves at one end of the distribution with relatively large scores; then come black and blue with decidedly smaller spans; and, finally, yellow is at the bottom with a very low apprehension score. The one distinguishable trend in the data is the slight negative correlation between luminosity of colors and apprehension score, i.e., the greater the luminosity the smaller the span. This is especially noticeable for the yellow letters which have the highest luminosity by far, and a span which is much smaller than those for any other colors. This trend is partly nullified, however, by the low apprehension scores for black and blue stimuli. These positions of black and blue, between yellow and the other colors, are inconsistent and apparently unexplainable in terms of luminosity, affective value, or attention value of the homogeneously colored letters.

There appears to be no relationship between apprehension scores in heterogeneous and homogeneous colors. Black, orange, and blue, which held positions 1, 2 and 3 in heterogeneous colors, had positions 6, 5, and 7, respectively, when the colored stimuli were homogeneously arranged. The only point of agreement is that yellow produced the lowest apprehension

score in both kinds of series. The correlation between ranking of homogeneous and heterogeneous colors is —.029. We noted above, however, that, if black and blue were omitted from the homogeneous series, there would be a high negative correlation (—.901) between span and relative luminosity of colors. In such a case there would be an appreciable correlation between the rankings of apprehension scores in the two kinds of series.

Consideration of the results from heterogeneously colored stimuli, however, reveals more consistent trends than were found with homogeneous colors. Both attention value and relative luminosity affect apprehension of heterogeneously colored letters to a considerable degree. The consistent trends in the results are undoubtedly influenced by the heterogeneity of the colors in the stimulus series. With eight different colors on each stimulus card the situation is more favorable for relative attention value to become effective than when each card has all one color which changes from card to card (successive series). In any heterogeneous series, therefore, a letter whose color has high attention value is apprehended more frequently than the letters whose colors possess less attention value.

In like manner, the relative luminosity of heterogeneously colored letters appears to have more effect than that of homogeneous colors on visual apprehension. With several colors of varying luminosity opportunity arises for letters differing in relative brightness to compete with each other in visibility due to the brightness contrast between colored letter and white background. It has been shown above that there is an appreciable

negative correlation between luminosity of colors and apprehension score. Apparently, this relation is caused by the difference in brightness contrast just described. This effect of relative luminosity is evidently different from attention value for no correlation exists between the two.

In a situation where letters of different colors are simultaneously exposed as in the heterogeneous color series, the various colors apparently exert a differential effect on apprehension (immediate memory) of the letters. It is possible that this factor also influences the total number of letters reproduced correctly after the exposure of any series. It will be remembered that in the homogeneous series this variation of color from letter to letter was absent. A comparison, therefore, of span for heterogeneous with that for homogeneous colors should reveal whatever influence heterogeneity of the colors has on total span. These comparisons are found in Table 8.

TABLE 8

EFFECT OF HOMOGENEOUS VERSUS HETEROGENEOUS COLORS ON SPAN OF VISUAL APPREHENSION

Scoring method	Homogeneous colors			Heterogeneous colors			D*	$\frac{D}{\sigma_D}$
	Group	Mean	σ_M	Group	Mean	σ_M		
I	100 M + W	5.75	.08	100 M	5.56	.09	+.19	1.63
I	100 M + W	5.75	.08	100 W	5.45	.09	+.30	2.55
II	100 M + W	6.22	.07	100 M	6.08	.08	+.14	1.23
II	100 M + W	6.22	.07	100 W	5.98	.08	+.24	2.18
III	100 M + W	5.27	.08	100 M	4.97	.09	+.30	2.35
III	100 M + W	5.27	.08	100 W	4.98	.10	+.29	2.19

*A plus indicates that the difference is in favor of the homogeneous colors.

As stated before, the observers for homogeneous colors were a mixed group with the same number of men and

women. For the heterogeneous colors, however, there were two subgroups, of 100 men, and 100 women, respectively. Before comparisons can be made between spans for the heterogeneous and homogeneous series, therefore, one must know whether sex differences occur in visual apprehension of colored letters. In Column 6 of Table 8 are found the mean scores of both men and women by the three methods of scoring. The average spans for men and for women were 5.56 and 5.45, respectively, in Scoring Method I; 6.08 and 5.98 in Method II; and 4.97 and 4.98 in Method III. The magnitude of the differences between these average scores of men and women are very small and statistically insignificant. For all practical purposes, the scores for men and women are approximately identical.

The comparison between span for homogeneous and heterogeneous colors is made by subtracting the mean score in Column 6 from the mean score in Column 3. The obtained difference is found in Column 8, and the ratio of the difference to the standard error of the difference in Column 9. A plus sign before any difference indicates that it is in favor of the homogeneously colored letters. Examination of the differences between all groups and by all methods of scoring reveals a range of +.14 to +.30, and shows that every difference is in favor of the homogeneous colors. While the direction of any one of these differences does not have great stability, the fact that all differences are in the same direction warrants the tentative conclusion that heterogeneity, in comparison with homogeneity of colored stimuli, tends to shorten the average span of visual apprehension to a slight degree.

IV

RELATED INVESTIGATIONS

There are a number of studies whose results have either a direct or an indirect bearing upon the influence of color on visual apprehension and perception. The purpose of the present chapter is to analyze these results in sufficient detail to yield, when correlated with the present investigation, a clear picture of the various factors operative in determining the apprehension and perception of both chromatic and achromatic stimuli.

Apprehension and Perception of Colored Stimuli

Hart (12) determined the range of visual attention, cognition, and apprehension[11] for colored stimuli. He calculated limens for each of four subjects in each of two series of stimuli. Red, blue, green, and yellow dots (Hering papers), pasted in homogeneous series on gray backgrounds, served as stimuli. The results for the four subjects have been combined into one group and the average limens ranked in Table 9. Results for all

TABLE 9

Ranks of Limens (1=Largest Limen) for Range of Attention, Cognition, and Apprehension of Colored Stimuli (from Hart)

	Attention			Cognition			Apprehension			All
Color	I	II	Both	I	II	Both	I	II	Both	three
Red	1	1	1	1	1	1	1	1	1	1
Blue	3	2	3	2	2	2	2	2	2	2
Green	4	4	4	4	4	4	4	3	3	4
Yellow	2	3	2	3	3	3	3	4	4	3

[11]Three typical processes, as determined from introspective reports, were first classified as "attention," "cognition," and "apprehension" by H. S. Oberly (18). Since each category involves apprehension, all are included in our summary table.

degrees of assurance are included (Hart's 5 + 4 + 3 data). The table should be read as follows: for the category "attention," in Series I, the ranks of the limens are: red = 1 (largest), blue = 3, green = 4, yellow = 2. Similarly for Series II, and both (I plus II) combined. The results may be stated in Hart's words: "In every case, red in general gives a high limen and green a low one. Yellow and blue have a tendency to be intermediate and variable. This situation is in general true for all 3 systematic categories and for all 3 degrees of subjective assurance . . ." While the individual results showed considerable variation, the summary of group results given in Table 9 reveals a rather marked tendency for the colors to take the ranks: red (1), blue, yellow, green.

Hart rejects intensity (luminosity) of the colors as an explanation of the differences obtained but suggests that differences in lag of visual sensation (*Anklingen* times) for the various colors is a more probable cause.

In considering the relative brightness of his colors, Hart has failed to evaluate the brightness of any color in relation to its background. The background was gray, but no statement concerning its brightness is given. It is probable that some of the colors (red and blue) were darker and the others (yellow and green) brighter than the background. It is also possible that red showed the greatest, green the least, and yellow and blue a moderate amount of brightness contrast between color and background.[12] If this be true, and the

[12]Percentage luminosity for standard Hering papers are approximately; red, 5.2; blue, 14.7; green, 44.3; and yellow, 75.7. If the neutral gray employed as background had 40 to 50 per cent of luminosity, the above statements are valid.

evidence is in its favor, then brightness or luminosity contrast between color and background does offer a satisfactory explanation of the differences between limens for apprehension of the colored stimuli. As will be shown later, retinal lag appears to be intimately related to intensity or brightness of the stimulus.

When one color is printed upon another as a background the legibility of the type may become very poor due to low visibility of the letters. Luckiesh (15, pp. 246-251) reports results for 13 combinations of colored print and background for printed matter read from a distance. The order from most to least legible follows:

1. Black	printed	matter	on yellow	background
2. Green	"	"	on white	"
3. Red	"	"	on white	"
4. Blue	"	"	on white	"
5. White	"	"	on blue	"
6. Black	"	"	on white	"
7. Yellow	"	"	on black	"
8. White	"	"	on red	"
9. White	"	"	on green	"
10. White	"	"	on black	"
11. Red	"	"	on yellow	"
12. Green	"	"	on red	"
13. Red	"	"	on green	"

Details of the experimental conditions such as kind of ink and paper employed, size of type, text used, etc., are omitted from the report. This is unfortunate because, when a white background is specified, one should know whether pure white, light gray, or light cream color is meant. Many kinds of print paper that are called white are really a light gray, and others are cream colored (light buff). Inspection of the above

list of color combinations, however, shows that the combinations involving the greater contrast between print and background, as black on yellow, blue on white, etc., are among the most legible. Those with small brightness contrast, as red on green and green on red, possess very poor visibility. The reason why black on white is in the middle of the series is not clear. This should have the greatest luminosity or brightness contrast. Perhaps the white background was not pure white, or a glare from the white background may have reduced the visibility of the black print. In general, however, Luckiesh's results show that legibility of the printed material is largely determined by the luminosity difference between color of print and color of background.

In a more carefully controlled experiment, Tinker and Paterson (29) studied the influence of variations in color of print and background on speed of reading. The Chapman-Cook Speed of Reading Tests, in which score on Form B is equal to score on Form A when typography is identical in the two forms, were utilized as the measuring instrument. Form A was printed with black ink on white Rainbow cover-stock and served as a standard in each comparison. Form B was printed with Ruxton's colored ink on Rainbow cover-stock in 10 variations of color combinations. Scores were obtained from 850 subjects. When score on Form B (color combination) was compared to score on Form A (black on white) and the color combinations ranked for influence on speed of reading, the order from most to least legible was found to be:

1.	Black	print	on white	background
2.	Green	"	on white	"
3.	Blue	"	on white	"
4.	Black	"	on yellow	"
5.	Red	"	on yellow	"
6.	Red	"	on white	"
7.	Green	"	on red	"
8.	Orange	"	on black	"
9.	Orange	"	on white	"
10.	Red	"	on green	"
11.	Black	"	on purple	"

Black on white, which has the greatest luminosity contrast between print and background, is distinctly the most legible text. Green on white, blue on white, and black on yellow, are all only slightly poorer than black on white. Red on yellow and red on white possess fair visibility, but all the remaining combinations (7 through 11) are poor. In the light of their results, Tinker and Paterson, in recommendations for hygienic printing, formulate the following rule: "In combining colors (color of ink and paper) care must be taken to produce a *printed page* which shows a maximum *brightness contrast* between print and background."

In the Tinker and Paterson experiment the measurement was in terms of speed of reading. To obtain a non-speed measure of the influence on legibility of variations in color of print and background, Preston, Schwankl, and Tinker (21) determined the effect of color combinations on the perception of isolated words. The greatest average distance from the eyes at which words printed in any color combination could be read constituted the measure of legibility for that combination. Black print on white paper was employed as a standard to compare with each of 10 other combina-

tions. The color combinations produced marked differences in comparative legibility of print. They are listed below, in order of legibility, the most legible combination in rank 1:

1.	Blue	print	on white	background
2.	Black	"	on yellow	"
3.	Green	"	on white	"
4.	Black	"	on white	"
5.	Green	"	on red	"
6.	Red	"	on yellow	"
7.	Red	"	on white	"
8.	Orange	"	on black	"
9.	Black	"	on purple	"
10.	Orange	"	on white	"
11.	Red	"	on green	"

These are the same color combinations used by Tinker and Paterson in the above-cited experiment. In both these studies, a more adequate interpretation of results is made possible by a knowledge of the appearance of the color combinations to an observer. The combinations, printed in Ruxton's ink on Rainbow cover-stock, are listed below with a description of the appearance of each in parenthesis after the trade name (color of backgrounds approached maximum saturation.):

Trade Name	*Observed Effect*
Black jobbing on white	(Black on light grayish white)
Grass green on white	(Dark green on light grayish white)
Lustre blue on white	(Dark blue on light grayish white)
Black jobbing on yellow	(Black on yellow with slight orange tinge)
Tulip red on yellow	(Light red on yellow with slight orange tinge)
Tulip red on white	(Light red on light grayish white)
Grass green on red	(Dark grayish green on red)
Chromium orange on black	(Dark lemon yellow on dark grayish black)
Chromium orange on white	(Light orange on light grayish white)

Tulip red on green	(Dark brown on dark green)
Black jobbing on purple	(Black on dark purple)

A comparison of the observed effect with the ranking of the above list for comparative legibility reveals the fact that, in general, the greater the luminosity difference between print and background, the greater the legibility of the color combination. The ranking for comparative legibility obtained in this study correlates +.864 with the ranking given by Tinker and Paterson who used a speed measure. This indicates that, whether measured in terms of speed of reading or perceptibility distance, the all-important factor conditioning perception of words is the brightness contrast between print and background.

With a somewhat different technique, Miyake, Dunlap, and Cureton (17) determined the relative legibility of black and colored numbers on colored and black backgrounds. In one series, single numbers were typed in black on red, green, yellow, and white backgrounds; and in a second series, red, green, yellow, and white numbers were printed on a black background. Fifteen subjects read the numbers from tachistoscopic exposure. The average number of times each color combination was read correctly determined the legibility ranks which follow, the most legible color combination receiving a rank of 1:

Series I

1. Black print on white background
2. Black " on yellow "
3. Black " on green "
4. Black " on red "

Series II

1. White print on black background
2. Yellow " on black "
3. Green " on black "
4. Red " on black "

These rank orders show that the combination of colored print and background which produced the greatest luminosity difference was found most legible, i.e., black on white in Series I, and white on black in Series II. In general, as the luminosity difference between printed numbers and background decreased, the legibility of the color combination decreased.

This survey and analysis of the investigations concerning the influence of color combinations on visual perception and apprehension in reading reveal a uniform trend common to all. In the first place, hue of the color apparently has little or no effect on perception or apprehension. The all-important factor seems to be the luminosity difference between the symbol to be apprehended and the background upon which it is printed. The data indicate this to be true for (1) the accuracy with which colored dots on a gray background are apprehended, (2) the speed with which material printed with colored ink on colored paper is read, (3) the distance at which words printed in colored ink on colored paper are perceived, and (4) the accuracy with which colored numbers on colored or white paper are apprehended.

Apprehension and Perception of Achromatic Stimuli

There have been several investigations on apprehension and perception of achromatic stimuli in which

variation in luminosity contrast between character and background occurred. Cooper (4), employing the systematic categories (attention, cognition, and apprehension) used by Hart (12), determined limens for black, dark gray, and light gray stimuli which consisted of paper dots pasted on a white background. The limens for the 4 subjects have been averaged and ranked in Table X. Results for all degrees of assurance are included (Cooper's 5 + 4 + 3 category). Under the category "attention," in Series I, black had the largest limen (rank 1), dark gray the next largest,

TABLE 10

RANKS OF LIMENS (1 = LARGEST LIMEN) FOR RANGE OF ATTENTION, COGNITION, AND APPREHENSION OF BLACK, DARK GRAY, AND LIGHT GRAY STIMULI (FROM COOPER)

	Attention			Cognition			Apprehension			All
Stimulus	I	II	Both	I	II	Both	I	II	Both	three
Black	1	1	1	1	2	1	1	1	1	1
Dark gray	2	3	2	3	3	3	2	2	2	2
Light gray	3	2	3	2	1	2	3	3	3	3

and light gray the smallest. The other columns are to be read in a similar manner. Examination of the group trends in this table reveals that black has the largest limen in all but one category (Series II, "cognition"). The ranks for dark gray and light gray vary considerably in "attention" and "cognition" limens, but consistently hold ranks 2 and 3, respectively, in "apprehension." The general trend for all three categories taken together (last column) shows black to hold rank 1, dark gray, rank 2, and light gray, rank 3. Cooper points out that stimulus intensities (brightness) do not appear to have any constant effect for the results of individual observers. In general, however, under the category

"apprehension" the darker stimuli produced larger limens, even with individual subjects. Summing the data of all subjects into one group accentuates the uniformity of results under the category "apprehension" and reveals a somewhat less consistent general trend running through all the data. The greater the brightness contrast between stimulus and background, the larger the limen, i.e., there is a negative correlation between luminosity of stimuli and size of limens, especially under category "apprehension."

Employing a somewhat different technique, Taylor and Tinker (25) determined the spans of visual apprehension for black, dark gray, and light gray letters on a white background. A total of 128 subjects observed in the experiment which was composed of a homogeneous series of letters in which all letters on any stimulus card were of the same brightness, as all black, etc., and a heterogeneous series in which each succeeding letter on any stimulus card varied in brightness, as black followed by dark gray, etc. The average scores of apprehension in each series follow:

Homogeneous Series		*Heterogeneous Series*	
Black	Av. = 24.63	Black	Av. = 25.24
Dark gray	Av. = 24.70	Dark gray	Av. = 24.50
Light gray	Av. = 23.88	Light gray	Av. = 21.06

Since the background was white, the greatest luminosity contrast between symbol and background occurred with black letters, next greatest with dark gray, and least with light gray. The above averages show that, in the homogeneous series black and dark gray yield approximately equivalent scores; but light

gray has a definitely smaller score. In the heterogeneous series more definite trends are noticeable. The average score for black is definitely the largest, dark gray next, and light gray the smallest by far. In both series the difference between light gray and either dark gray or black has rather high statistical significance, especially in the heterogeneous series. In general, the results indicate that the greater the luminosity difference between symbol and background, the greater the average score of apprehension.

Ferree and Rand (8), in their series of experiments on intensity of light and speed of vision, have investigated the comparative effects for dark objects on light backgrounds and light objects on dark backgrounds. Broken circles subtending visual angles of 1 to 3 minutes were used as test objects. Eleven degrees of illumination intensity, from 1.25 to 100 foot-candles, were employed. They determined the shortest exposure during which a symbol could be distinguished accurately. The shorter this time (faster the vision), the more legible the symbol for any given illumination intensity and size of test object. Four combinations of brightness were used in constructing the test objects: black on white, white on black, white on gray, and black on gray. Below are listed the brightness combinations from fastest (most legible) to slowest speed of vision for each size of test object:

Visual angle equals 1 minute

1.25 to 5.00 foot-candles: black on white fastest, white on black next, white on gray slowest.

7.50 to 100 foot-candles: white on black fastest, black on white next, white on gray next, black on gray slowest.

Visual angle equals 2 minutes

1.25 to 100 foot-candles: white on black fastest, black on white next, white on gray next, black on gray slowest.

Visual angle equals 3 minutes

1.25 foot-candles: white on black fastest, white on gray next, black on white next, black on gray slowest.

2.50 to 10.00 foot-candles: white on black fastest, black on white next, white on gray next, black on gray slowest.

15.00 to 100 foot-candles: white on black fastest, white on gray next, black on white next, black on gray slowest.

In considering the above results the following statements of Ferree and Rand should be kept in mind: "There is a greater difference in sensation between object and background in case of white on black than black on white, due probably to physiological induction or contrast" when determinations of speed of discrimination are made well above the threshold of acuity. "The gray background used is nearer to the black than to the white test-letter in sensation value. There is, therefore, greater contrast between object and background in case of the white than the black test-letter."

In general, the above rank orders of brightness combinations for speed of vision take the positions; white on black fastest, black on white next, white on gray next, and black on gray slowest.[18] This rank order corresponds exactly with the magnitude of luminosity difference between symbols and backgrounds. The greater the luminosity contrast between test-object and background, therefore, the faster the speed of vision (greater the legibility).

---

[18]The fact that black on white has a faster speed than white on black for low intensities with small test objects is explained by the authors in terms of the nature of the test object.

In a related study, which may be interpreted in the same way as the above, Kirschmann (14) found that both block letters and geometric figures were more easily read in white print on black background than in black on white. The measure of legibility was the angular distance from the fixation point at which the letters and figures were recognized correctly in indirect vision.

Starch (22, pp. 668-669) reports an experiment in which the problem of brightness contrast is studied in a normal reading situation. Two pieces of text were set up exactly alike, except that one was printed in black type on a white background, and the other in white type on a dark gray background. These selections were read by 40 subjects at their natural rate. The average number of words read per second were:

Black on white	6.06
White on dark gray	4.26

These results show a difference of 42 per cent in favor of the black type on white background. The text showing the greater luminosity contrast between print and background yielded the fastest reading.

In a somewhat similar experiment, Paterson and Tinker (19), using the two equivalent forms of the Chapman-Cook Speed of Reading Tests with 280 subjects, compared the speed of reading black print on white background with white print on black background. They found a 10.5 per cent difference in favor of the black on white printing arrangement.

Holmes (13), employing, as a measure of legibility, the distance from the eye at which a symbol could be

read correctly, compared the perceptibility of isolated words printed in black type on white background with words printed in white type on black background. Averages from 20 subjects showed a 14.7 per cent advantage for the black type on a white background. The results of the last two investigations appear to contradict the previously cited findings of Ferree and Rand. There is urgent need of further study of the legibility of black versus white print *in the normal reading situation*. There are suggestions of factors other than brightness contrast, which may cause the differences in favor of the black on white in perceiving printed words. A study, now in progress at Minnesota, is making an analysis of the influence of size and nature of the printed symbols, and other factors which may influence the perceptibility of black and white print.

This summary and analysis of the experimental results dealing with effect of brightness combinations on visual apprehension and perception in reading show an almost universal trend. The most important factor conditioning perceptibility of printed characters appears *to be the luminosity difference between the* symbol to be apprehended and its background. The cited results show this to be true for: (1) the accuracy with which black and gray dots on a white background were apprehended; (2) the accuracy with which black and gray letters on a white background were apprehended; (3) the speed with which text printed in black on white and white on gray was read; and (4) the speed with which a test object in black on white, white on black, black on gray, and white on gray was dis-

criminated. The few exceptions appear to be adequately explained as due to other factors such as nature of the test object or printed character, size of printed symbol, or intensity of illumination.

Lag of Visual Sensation

Several investigations have concerned themselves with the influence of brightness (luminosity) and color on the lag of visual sensation. Data reported in these experiments yield indirect evidence concerning the influence of difference in luminosity contrast between symbol and background on visual apprehension.

In one of the most carefully controlled experiments in this field, Bills (3) made determinations by previously used methods as well as by an improved method devised at the Bryn Mawr laboratory. In her own study she determined the time in seconds required for sensation to rise to its maximum value when the stimuli (red, yellow, green, blue, and white lights) were all photometrically equal; and also when different luminosities of the same color were employed. In Bills' experimental situation, the lights to be judged appeared on and completely covered a small white surface. This white surface rested on a black velvet background. The intensity of the light stimulus, therefore, was the same as the amount of contrast between the white surface and the black background. Any change in intensity of the lights also produced a change in luminosity contrast between the illuminated surface and its background.

Bills' results, by Method 3, are summarized in Table 11. The lights employed to produce the sensations

under investigation were made constant at either .057, or .150, or 1.21 meter-candles. The durations of visual

TABLE 11
TIME IN SECONDS REQUIRED FOR SENSATION TO RISE TO ITS MAXIMUM VALUE (FROM BILLS, METHOD 3)

	Time of lag when intensity of light in meter-candles is constant at:		
Color used	.057	1.50	1.21
Red	0.164	0.148	0.100
Yellow	0.103	0.086	0.088
Green	0.190	0.146	0.136
Blue	0.210	0.134	———
White	0.216	0.184	0.105

lag are given in Columns 2, 3, and 4. For example, it took the red sensation with a photometric value of .057 meter-candles 0.164 seconds to rise to its maximum value; of .151 meter-candles, 0.148 seconds; and of 1.21 meter-candles, 0.100 seconds. The data in this table reveal an almost universal tendency for an increase in brightness (brightness = brightness contrast) to increase the speed with which a sensation rises to its maximum value. The single exception is yellow, whose visual lag is approximately the same for 1.21 and .150 meter-candles. "With increase of intensity of light," therefore, "there was a decrease in the time required to produce the maximum response."

Many other studies cited by Bills yield abundant evidence which supports her findings. Exner, Lough, Martius, Broca and Sulzer, McDougall, and Buchner[14] all found that the time required for visual sensation to reach its maximum value varied inversely with

[14]See Bills (3) for a comprehensive review of these experimental results.

the intensity of the stimulus light. The results of McDougall (16) are typical and are listed below:

Intensity values	Sensation lag
1 unit	0.049 second
1/2	0.055
1/4	0.061
1/8	0.066
1/16	0.078
1/32	0.089
1/64	0.100
1/128	0.127
1/256	0.142
1/512	0.150
1/1024	0.183
1/2048	0.200

These results of McDougall reveal a constant increase in duration of visual lag with decrease in intensity of white sensation.

The consistency of the trend of evidence in these experiments justifies the conclusion that the greater the brightness contrast between a spot of light and its background, the shorter the time required for sensation to reach its maximum value. This principle of visual lag appears to be, therefore, one of the important determinants of visual apprehension. It must be considered as one of the more important factors which help to explain the effect of brightness difference between figure (symbol) and background on visual apprehension, speed of reading, perceptibility of words, and other similar perceptual processes.

While the influence of brightness of stimulus on visual lag appears to be conclusively demonstrated, the evidence concerning effect of hue of color on lag is not at all final. Whenever the colored stimuli used are not equated for brightness, experimental results show that the brighter colors yield least sensation lag. If the

colors are photometrically equal, however, varying results have been reported. McDougall, for example, found that colored lights of equal photometric value all required the same time to produce their maximum effect of sensation. Bills, however, with lights photometrically equal at each of three different intensities, found that the times required for various colored sensations to rise to their maximum value were all different. A summary of her results by method 3 are given in Table 11. At present, therefore, one cannot state with any degree of assurance that hue of color has a consistent influence on visual lag.

Evidence from the various sources cited indicates that differences in visual apprehension and perception of both colored and achromatic symbols are probably due, *at least in part, to lag of visual sensation which is* produced by variation in luminosity contrast between character to be apprehended and background.

Both direct and indirect evidence from a number of related investigations warrant the following conclusions concerning comparative potency of hue and luminosity of color on visual apprehension and perception of symbols: (*a*) Hue of color has little or no effect on apprehension and perception. (*b*) The luminosity contrast between symbol to be apprehended and its background has a large and very important influence on apprehension and perception. (*c*) Lag of visual sensation, which is due to brightness contrast between symbol and background, probably explains to a large degree the differences obtained in visual apprehension and perception of symbols varying in color and luminosity.

V

SUMMARY AND CONCLUSIONS

Because color and color combinations are widely employed in situations designed to convey messages by means of visual symbols, there is need of adequate information concerning the influence of color on visual apprehension and perception. Whenever colored letters, words, or other symbols are used on either a white or a colored background in printing advertisements, constructing automobile license plates, and the like, there is danger that the words or other symbols may lack adequate visibility. Other things being equal, the color combinations which favor quick and accurate apprehension of printed characters in all perceptual situations should be chosen. With a knowledge of the comparative legibility of symbols involving various color combinations, the printer will be able to make use of colors to obtain affective value or attention value, and, at the same time, maintain adequate visibility of textual material.

In this investigation the effect of color on visual apprehension and perception has been studied by an analysis of the apprehension scores for (1) homogeneously colored letters in which all symbols on any stimulus card were of the same color, and (2) heterogeneously colored letters in which each succeeding letter on any stimulus card was of a different color. The influence of affective value, attention value, and luminosity of colors on apprehension and perception of symbols were included in the analysis. Other features

of the experiment included determinations of (1) the effect of scoring methods on average span of visual apprehension and on reliability of scores, and (2) the influence of letter position on visual apprehension.

There were eight letters on each stimulus card, and 32 cards in the homogeneous (4 of each color) and 24 cards in the heterogeneous series. The eight colors employed were: black, orange, violet, blue, red, neutral gray, green, and yellow. A group of 100 university students (50 men plus 50 women) were subjects for the experiment with homogeneous colors, and 100 university men and 100 university women were subjects for the investigation involving heterogeneous colors. The exposure interval was three seconds for all series. Color preferences were obtained for the eight colors by the method of paired comparisons.

To determine the influence of scoring method on span of visual apprehension the responses were scored in three ways: (1) average span computed from scores in which a credit of one was given for each item reproduced correctly and in the right place, and in addition a credit of one-half was given for each item reproduced correctly but out of place in the series (Method I); (2) average span calculated from scores in which each item correctly reproduced received a credit of one, irrespective of whether it was or was not in the right place (Method II); (3) average span derived from scores in which a credit of one was given for each item correctly reproduced and in the right place, but no credit for items reproduced correctly but in a wrong position (Method III).

Scoring Method III yielded the smallest average span of 4.98 letters; Method I came next with a score of 5.45; and Method II produced the largest span which was 5.98. The directions of the differences between averages are very stable. The difference between either I and II or II and III is approximately one-half an item; and between I and III, about one item on the average.

Scores in Method I correlated very high with both II and III, the coefficients being .920 and .973, respectively. The correlation of .916 between II and III is nearly as high.

These results justify the conclusion that, when it is desirable to know absolute span of apprehension, method of scoring is important. If only relative span (position of individual span in the group) is wanted, however, any one of the three methods of scoring may be employed, with a slight preference for Method I.

Reliability was computed by correlating the sums of the odd versus the sums of the even scores and then applying the Brown-Spearman "prophecy" formula. The raw coefficients ranged from .786 to .908 in Method I; from .798 to .866 in Method II; and from .768 to .819 in Method III. Method I and Method II, therefore, have approximately the same reliability, which is slightly higher than that of Method III. All reliability coefficients are relatively high, which indicates that all three methods of scoring yield results with high internal consistency. Although Method I appears to be slightly more satisfactory, any one of the three scoring methods may be employed and still

achieve an adequate measure of visual apprehension.

Letter position in a stimulus series was found to have a definite effect on visual apprehension. From left to right there was a decrease in the average number of letters correctly reproduced in each succeeding letter position through the seventh, and then a slight increase in score at the last position. The decrease from the first to the fourth position was constant and gradual. There were rapid drops in score from the fourth to fifth and from sixth to seventh positions. These rapid drops together with the increase in score at position eight produced marked irregularities of trend from position to position in the series.

In the homogeneous color series neither affective value nor attention value of colors influenced apprehension of letters. Luminosity of colors, however, had some effect on apprehension. In general, the greater the luminosity of the colored letters, the smaller the perceptual span. A marked example of this tendency was the consistently low score for yellow letters. Two striking exceptions to this generalization were the low spans for black and blue, both of which have a low percentage of luminosity. The correlation between luminosity and apprehension score is —.351, but this becomes —.901 if the results for black and blue are omitted.

The results for the heterogeneous color series revealed more consistent trends. There appeared to be only slight sex differences in apprehending colored letters or in color preferences. In the analysis of the differences in scores for the various colors, affective

value of color showed no correlation with apprehension score. The greater the attention value, however, the greater the apprehension score for that color since attention value correlated +.524 with apprehension score. Luminosity showed a definite effect on apprehension of heterogeneously colored letters. The correlation of —.667 shows that, in general, the greater the luminosity the smaller the score. These effects of attention value and the relative luminosity are apparently independent of each other, for their intercorrelation is —.130.

The heterogeneous arrangement of colored letters permitted the influence of luminosity on visual apprehension to become more definite and consistent than with homogeneous colors. Since all colored letters were on a white background, the general trend in both arrangements of stimuli was for a larger apprehension score to occur with the greater luminosity contrasts between color and background.

Span of visual apprehension for colored letters in the heterogeneous arrangement of stimuli was slightly smaller than in the homogeneous series.

There are three types of related investigations: (1) apprehension and perception of colored stimuli; (2) apprehension and perception of achromatic stimuli; and (3) lag of visual sensation.

A survey of the first group of experiments revealed that hue of color apparently has little or no effect on apprehension and perception. The important factor influencing apprehension and perception in reading seemed to be the luminosity difference between the

symbol to be apprehended and the background upon which it was printed. This was found to be true for (1) the accuracy with which colored dots on a gray background were apprehended; (2) the speed with which material printed with colored ink on colored paper was read; (3) the distance at which words printed in colored ink on colored paper were perceived; and (4) the accuracy with which colored numbers on colored and white paper were apprehended.

The summary and analysis of studies dealing with the effect of brightness combinations on visual apprehension and perception in reading showed again that the *luminosity difference between the symbol to be* apprehended and the background upon which it was printed was an important determinant of apprehension. This was found *to hold for* (1) the accuracy with which black and gray dots and letters on white backgrounds were apprehended; (2) the speed with which text printed in black on white and white on gray was read; and (3) the speed with which a test object in black on white, white on black, black on gray and white on gray was discriminated.

Evidence from several experimental studies indicates that differences in visual apprehension and perception of both colored and achromatic symbols are probably due, at least in part, to lag of visual sensation which is produced by variation in luminosity contrast between character to be apprehended and background upon which it is printed.

Both direct and indirect evidence from a number of related experiments warrants the following conclusions

concerning the comparative potency of hue and luminosity of colors on visual apprehension and perception of symbols: (1) Hue of color has little or no effect on apprehension and perception except with heterogeneous color series in which attention value of colors is one important determinant of apprehension. (2) The luminosity contrast between symbol and background has the greatest influence on apprehension and perception. (3) Lag of visual sensation, which is due to brightness contrast between symbol and background, probably explains to a large degree the differences obtained in visual apprehension and perception of symbols varying in color and luminosity.

Results obtained in the present experiment and those reported in related investigations show close agreement. The all-important determinant of visual apprehension and perception of printed words, letters, and similar symbols is the luminosity contrast between character and background.

These results have a direct bearing upon special printing situations such as advertising and posters where colors are employed for attention value, affective value, and the like. In any situation of this kind where quickness of perception is essential, care should be taken to use a color or brightness combination which produces a maximum brightness contrast between symbol and background.

References

1. Adams, H. F. Advertising and its mental laws. New York: Macmillan, 1916. Pp. xi+333.
2. Anderson, A. C., & Bolton, F. J. Inhibition of the unpleasant. *J. Abn. & Soc. Psychol.*, 1925, **20**, 300-302.
3. Bills, M. A. The lag of visual sensation in its relation to wave lengths and intensity of light. *Psychol. Monog.*, 1920, **28**, No. 127. Pp. 101.
4. Cooper, S. F. The effect of brightness in the range of attention experiment. *Amer. J. Psychol.*, 1928, **40**, 254-274.
5. Crosland, H. R. The influence of letter position on range of apprehension—a reply to Dr. Tinker. *Psychol. Bull.*, 1929, **26**, 375-377.
6. Crosland, H. R., & Johnson, G. The range of apprehension as affected by inter-letter hair spacing and by the characteristics of individual letters. *J. Appl. Psychol.*, 1928, **12**, 82-124.
7. DeCamp, J. E. The influence of color on apparent weight. *J. Exper. Psychol.*, 1917, **2**, 347-370.
8. Ferree, C. E., & Rand, G. Intensity of light and speed of vision: II. Comparative effects for dark objects on light backgrounds and light objects on dark backgrounds. *J. Exper. Psychol.*, 1930, **13**, 388-422.
9. Gordon, K. Ueber das Gedächtnis für affektiv bestimmte Eindrücke. *Arch. f. d. ges. Psychol.*, 1905, **4**, 437-458.
10. ————. The recollection of pleasant and unpleasant odors. *J. Exper. Psychol.*, 1925, **8**, 225-239.
11. Guilford, J. P., & Dallenbach, K. M. The determination of memory span by the method of constant stimuli. *Amer. J. Psychol.*, 1925, **36**, 621-628.
12. Hart, J. R. The range of visual attention, cognition and apprehension for colored stimuli. *Amer. J. Psychol.*, 1928, **40**, 275-283.
13. Holmes, G. The relative legibility of black print and white print. *J. Appl. Psychol.*, 1931, **15**, 248-251.
14. Kirschmann, A. Ueber die Erkennbarkeit geometrischer Figuren und Schriftzeichen im indirekten Sehen. *Arch. f. d. ges. Psychol.*, 1908, **13**, 352-388.
15. Luckiesh, M. Light and color in advertising and merchandising. New York: Nostrand, 1923. Pp. 246-251.

16. McDougall, W. The variation of the intensity of visual sensations with the duration of the stimulus. *Brit. J. Psychol.*, 1904, 1, 151-189.

17. Miyake, M. F., Dunlap, J. W., & Cureton, E. E. The comparative legibility of black and colored numbers on colored and black backgrounds. *J. Gen. Psychol.*, 1930, 3, 340-343.

18. Oberly, H. S. The range for visual attention, cognition, and apprehension. *Amer. J. Psychol.*, 1924, 35, 332-352.

19. Paterson, D. G., & Tinker, M. A. Studies of typographical factors influencing speed of reading: VI. Black type versus white type. *J. Appl. Psychol.*, 1931, 15, 241-247.

20. Poffenberger, A. T. Psychology in advertising. New York: Shaw, 1928. Pp. xix+632

21. Preston, K., Schwankl, H. P., & Tinker, M. A. The effect of variations in color of print and background on legibility. *J. Gen. Psychol.*, 1932 (in press).

22. Starch, D. Advertising. New York: Scott, Foresman, 1914 Pp. 189-190.

23. ———. Principles of advertising. Chicago: Shaw, 1923 Pp. 668-669

24. Tait, W. D. The effect of psychophysical attitudes on memory. *J. Abn. Psychol.*, 1913, 8, 10-37.

25. Taylor, C. D., & Tinker, M. A. The effect of luminosit on the apprehension of achromatic stimuli. *J. Gen. Psychol* 1932 (in press).

26. Tinker, M. A. Visual apprehension and perception in reac ing. *Psychol. Bull.*, 1929, 26, 223-240.

27. ———. The influence of letter position on range of visu apprehension—a reply to Dr. Crosland. *Psychol. Bull.*, 192 26, 611-613.

28. ———. Physiological psychology of reading. *Psycho Bull.*, 1931, 28, 81-98.

29. Tinker, M. A., & Paterson, D. G. Studies on typographic factors influencing speed of reading: VII. Variations in col of print and background. *J. Appl. Psychol.*, 1931, 15 (press).

L'EFFET DE LA COULEUR SUR L'APPRÉHENSION ET LA PERCEPTION VISUELLES

(Résumé)

Dans la partie principale de cette enquête, on a étudié l'effet de la couleur sur l'appréhension et la perception visuelles au moyen d'une analyse des résultats de l'appréhension (1) au cas des lettres colorées homogènes où tous les symboles sur une carte quelconque servant de stimulus ont été de la même couleur, et (2) au cas des lettres colorées hétérogènes où chaque lettre suivante sur une carte quelconque servant de stimulus a été d'une couleur différente. D'autres parties de l'expérience ont compris les déterminations (1) de l'effet des méthodes d'évaluation sur l'étendue moyenne de l'appréhension visuelle et sur la constance des résultats, et (2) de l'influence de la position des lettres sur l'appréhension visuelle. Pour déterminer l'influence de la méthode d'évaluation sur l'étendue de l'appréhension visuelle, on a évalué les réponses selon trois méthodes. La Méthode d'évaluation III, où l'on a donné des points pour des parties reproduites en l'ordre qu'il faut, a montré la plus petite étendue moyenne de l'appréhension; la Méthode I, où l'on a donné un demi-point de plus pour chaque partie reproduite en l'ordre incorrect, a eu une étendue moins petite; et la Méthode II, où l'on donné des points pour toutes les parties reproduites sans égard de l'ordre, a eu la plus grande étendue. La différence entre la Méthode I et la Méthode II, ou entre les Méthodes II et III, a été d'à peu près une demi-partie; et entre I et II environ une partie en moyenne. Toutes les méthodes ont montré une haute corrélation l'une avec l'autre. On conclut que la méthode d'évaluation a un effet important sur l'étendue absolue mais peu d'influence sur l'étendue relative de l'appréhension visuelle. Toutes les méthodes d'évaluation ont eu une constance relativement haute; mais la constance de la Méthode III a été un peu moins haute que les constances des Méthodes I et II, lesquelles ont été approximativement égales l'une à l'autre. On a trouvé que la position des lettres dans une série servant de stimulus a *un effet défini sur l'appréhension visuelle. De gauche à droite* il y a eu une décroissance du nombre moyen de lettres correctement reproduites dans chaque position suivante des lettres, jusqu'à la septième, et puis une petite croissance à la dernière position. Dans les séries de couleurs homogènes ni la valeur affective ni la valeur des couleurs pour l'attention n'ont influé sur l'appréhension des lettres. Plus la luminosité des lettres colorées est grande, cependant, plus l'étendue de la perception est petite. Dans les séries de couleurs hétérogènes la valeur affective des couleurs n'a montré aucune corrélation avec les résultats de l'appréhension; mais la valeur pour l'attention a donné une corrélation positive et la luminosité une corrélation négative avec les résultats de l'appréhension dans cette disposition des couleurs. L'évidence de plusieurs études expérimentales indiquent que les différences dans l'appréhension et la perception visuelles des symboles colorés et achromatiques seraient dues, du moins en partie, au retard de la sensation visuelle lequel est produit par la variation dans le contraste de la luminosité entre la lettre à appréhender et le fond sur lequel elle est imprimée. Les résultats obtenus dans cette expérience et ceux rapportés dans les enquêtes pareilles sont très en accord. Le déterminant le plus important de l'appréhension et de la perception visuelles des mots, des lettres, et des symboles semblables imprimés est le contraste de luminosité entre le symbole et le fond.

TINKER

DIE EINWIRKUNG DER FARBE AUF DIE VISUELLE AUFFASSUNG UND WAHRNEHMUNG

(Referat)

In dem Hauptteil dieser Untersuchung erforschte man die Einwirkung der Farbe auf die visuelle Auffassung und Wahrnehmung, mittels einer Analyse der erzielten Auffassungszahlen (apprehension scores) bei 1) homogen gefärbten Buchstaben, wobei alle Symbole auf jeder besonderen Reizkarte (stimulus card) die selbe Farbe hatten, und 2) bei heterogen gefärbten Buchstaben, wobei jede der aufeinanderfolgenden Buchstaben auf einer gegebenen Reizkarte ihre besondere Farbe hatte. Unter den Bestandteilen der Untersuchung fanden sich auch 1) Bestimmungen der Einwirkung der Zahlberechnungsmethoden (scoring methods) auf den mittleren Wert der visuellen Auffassungsspannweite (visual apprehension span) und auf die Zuverlässigkeit der Zahlen, und 2) Bestimmungen der Einwirkung der Stellung (position) der Buchstabe auf die visuelle Auffassung. Um die Einwirkung der Zahlberechnungsmethode auf den mittleren Wert der visuellen Auffassungsspannweite zu bestimmen, berechnete man die Reaktionen (responses) auf drei Weisen. Berechnungsmethode Nummer III, womit der Vp. nur diejenigen Buchstaben zugerechnet wurden, die in der richtigen Anordnung wiederholt worden waren, ergab den kleinsten mittleren Wert der Auffassungsspannweite; die Methode I, womit für jede Buchstabe die in falscher Anordnung wiederholt wurde doch einen halben Punkt hinzugerechnet wurde, ergab die nächst-kleinste Spannweite; und die Methode II, womit der Vp. alle wiederholte Buchstaben zugerechnet wurden, ohne Rücksicht auf die Anordnung, ergab die grösste Spannweite. Der Unterschied zwischen den Methoden I und II oder den Methoden II und III betrug im Durchschnitt ungefähr einen halben Punkt, und zwischen I und III ungefähr einen Punkt. Es bestand eine hohe Korrelation unter allen Methoden. Man schloss hieraus, dass die Berechnungsmethode zwar auf die absolute Spannweite der visuellen Auffassung eine starke, auf die relative Spannweite aber nur eine schwache Einwirkung hatte. Alle Berechnungsmethoden erwiesen eine hohe Zuverlässigkeit, aber die Zuverlässigkeit der Methode III war etwas geringer als die Zuverlässigkeiten der Methoden I und II, die fast gleich waren. Es zeigte sich, dass die Anordnung der Buchstaben in einer Reizserie eine bestimmte Einwirkung auf die visuelle Auffassung ausübte. Von links nach rechts zeigte sich eine Verminderung der mittleren Zahl der richtig wiederholten Buchstaben in jeder sukzessiven Buchstabenstellung bis durch die siebente, und dann in der letzten Stellung eine geringe Erhöherung der Zahl.

Bei der homogenen Farbenserie beeinflusste weder der affektive Wert noch der Aufmerksamkeitswert (attention value) der Farben die Auffassung von Buchstaben. Je stärker, jedoch, der Glanz (luminosity) der farbigen Buchstaben, desto kleiner war die Spannweite. Bei den heterogenen Farbenserien zeigten affektive Werte ebenfalls keine Beziehung zur Auffassungszahl. Bei dieser Anordnung der Farben ergab aber der Aufmerksamkeitswert eine positive und der Glanz eine negative Korrelation mit der Auffassungszahl. Beweis aus mehreren experimentellen Untersuchungen deutet darauf hin, dass die Unterschiede bei der visuellen Auffassung und Wahrnehmung sowohl von farbigen wie von achromatischen Symbolen wahrscheinlich wenigstens teilweise auf eine Verzögerung der visuellen Empfindung zurückzuführen ist,—eine Verzögerung die durch Variation des Glanzkontrastes zwischen der aufzufassenden Buchstabe und dem Hintergrund, worauf sie gedruckt ist, verursacht wird. Es besteht eine ziemlich

genaue Uebereinstimmung zwischen den Befunden aus dieser Untersuchung und den Befunden, die aus verwandten Untersuchungen gemeldet worden sind. Der äusserst-wichtige Faktor bei der Bestimmung der visuellen Auffassung und Wahrnehmung gedruckter Wörter, Buchstaben, und ähnlicher Symbole ist der Glanzkontrast zwischen Symbol und Hintergrund.

TINKER

$7.00 per volume
Single numbers $[illegible]

MONTHLY
Two volumes per year

March, 1932
Volume XI, No. 3

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

EDITED BY
CARL MURCHISON

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the post-office at Worcester, Mass., under Act of March 3, 1879

$7.00 per volume
Single numbers $2.00

MONTHLY
Two volumes per year

March, 1932
Volume XI, No. 3

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

THE RELIABILITY AND VALIDITY OF MAZE EXPERIMENTS WITH WHITE RATS*

From the Psychological Laboratories of Clark University

By

ROBERT LEEPER

*Accepted for publication by Walter S. Hunter of the Editorial Board and received in the Editorial Office January 6, 1931.

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the postoffice at Worcester, Mass., under Act of March 3, 1879

CONTENTS

I. INTRODUCTION 141

II. HISTORICAL REVIEW 144

III. THE LOGICAL AND STATISTICAL PROBLEMS OF RELIABILITY AND VALIDITY AS RELATED TO MAZE EXPERIMENTS 151

- A. The Formulae for the Reliability of Differences between Group Averages . . . 151
- B. Problems of Validity 159
- C. Problems of Reliability Coefficients . . 161

IV. EXPERIMENTAL RESULTS 174

- A. Introduction 174
- B. Apparatus and Method 175
  - 1. The Straightaway Used for Preliminary Training 175
  - 2. Construction of the Maze 175
  - 3. The Room Used to Control Sound . 178
  - 4. Animals 180
  - 5. Procedure 183
- C. Results 187
  - 1. Performance on the Straightaway During the Preliminary Training . . 189
  - 2. Learning Curves of the Different Groups 191
  - 3. Distribution of Scores in Different Periods of Learning 198
  - 4. Correlations between Scores on the First Trial and Scores on Subsequent Trials 203
  - 5. Relation of Weight Changes to Learning Curves and Variability 205
  - 6. Reliability Coefficients from the Present Experiment 211
    - *a.* Correlations of Odd *vs.* Even Trials 213
    - *b.* Correlations of Groups of Ten Trials and of Fifteen Trials . . 215
    - *c.* Correlations from Groups of Three Trials 218
    - *d.* Correlations of Scores on Different Mazes 224
    - *e.* Correlations of Test and Retest . 228

V. GENERAL SUMMARY AND CONCLUSION . . . 230

REFERENCES 240

RÉSUMÉ EN FRANÇAIS 243

REFERAT AUF DEUTSCH 244

I

INTRODUCTION

The experimental work reported in this paper deals exclusively with the problem of reliability; the theoretical discussion is concerned with both reliability and validity. Both of these problems are of fundamental importance for a science of animal behavior—knowledge of validity to enable investigators properly to interpret whatever experimental results are secured, and knowledge of reliability to enable investigators to get results sufficiently free from the operations of chance so that the data can illustrate whatever principles are at work.

From certain standpoints it may appear that detailed quantitative studies such as this are premature. It is probably quite true, as Köhler maintains, that psychology is so young a science that there is still considerable need for exploratory experiments of a qualitative type and that any attempt now in many fields to secure great precision of measurement would result only in the calibration of experimental procedures with relatively slight potentialities for scientific research. It may even prove that, for the study of learning in animals, maze experiments may be quite unsatisfactory in comparison with some other methods such as that of conditioned reflex experiments. However, several considerations encourage the student of the reliability of maze experiments. In the first place, various investigations such as Tryon's study (33) of the inheritance of learning ability, Maurer and Tsai's study (19) of the influence

of Vitamin B deficiency during the nursing period on later learning ability have seemed to indicate that there are certain valid purposes which maze experiments can serve, such as the study of heredity, the effect of drugs on learning, etc.,—all of which depend upon quantitative comparisons. It remains to be demonstrated that maze learning is less suited for such studies than other types of learning experiments. In the second place, development of statistical practices in the field of maze experiments should be of advantage in other fields of learning study as well. In a sense, the fundamental statistical devices used here are not different from those developed relative to the problems of the reliability of various tests and measurements, but the details of application introduce a number of new and puzzling problems.

The purpose of the present study, then, is to throw light on the problem of how to secure dependable quantitative results in studies of learning, and especially in studies of maze learning with white rats. As a means to this end, the aim has been, first, to make an analysis of the logical and statistical problems of reliability and validity as they apply to maze experiments, and, secondly, to make a critical experimental examination of the reliability of experiments using the multiple-T maze, particularly with respect to the effects of permitting various degrees of retracing and with respect to the feeding program used. In seeking to discover the reliability of this maze, reliability coefficients have been calculated by test and retest [a method recently used by Heron (6) under one particular arrangement, and

by Tryon (30) under another particular condition], as well as by several of the methods commonly used to measure the consistency of performance of rats within the training on a single maze.

There is not attempted in the theoretical section of this paper a complete discussion of the logical and statistical problems of reliability and validity, although such a discussion would not be out of place, since the past discussions of the reliability of maze experiments have often been very loosely conducted. The recent article by Tolman and Nyswander (28), however, has furnished an excellent critical review of many phases of the problem, and on these particular points I would only echo what they have already said.

In the experimental work of this study and in the methods of treating experimental data, this study is closely related to those by Tolman and Nyswander (28), Stone and Nyswander (25), and Tryon (30). In its consideration of the theoretical problems of the reliability and validity of maze experiments, it depends not only on the work of those who have contributed to this problem directly (especially Heron, Hunter, Stone, Tolman, and Tryon), but also on the developments of the problems of reliability and validity in the fields of intelligence tests and educational tests, where these problems have received a much earlier and more mature treatment.

The study was conducted at Clark University in 1928-1930 under the direction of Dr. W. S. Hunter, and with advice and suggestions from Dr. Vernon Jones and Dr. R. R. Willoughby on a number of statistical points.

II

HISTORICAL REVIEW

The statistical study of the reliability of experiments on animal behavior may be said virtually to have started in 1917 with Paterson's review (22) of three articles by Bassett, Hubbert, and Ulrich. Paterson showed that, in spite of the quite positive conclusions all three experimenters had drawn, only Ulrich's conclusions seemed justified when the reliability of the group differences was examined in terms of the ratio of the difference to the probable error of the difference. Since the publication of Paterson's article (though, of course, not primarily because of it, but because of the development of statistical procedures in other fields of psychology) an increasing proportion of studies of animal behavior have availed themselves of the statistical measures of reliability. It is still true, however, that too many studies are to some extent vitiated by the neglect of statistical treatment.

One indirect suggestion in Paterson's article was that the maze procedures used might be too unreliable to permit satisfactory measurement of any ordinary experimental effects, and that, consequently, the first maze problem to attack was that of developing more reliable methods in maze experiments. Paterson himself began such experiments, but the results were not published. With Paterson's permission, Hunter and his students then took up the problem. The data from these early experiments showed very low reliabilities for current maze techniques, and served to awaken animal experi-

menters to a realization of the fact that with the type of maze pattern and maze method in general use at that time there could be little hope of ever securing dependable results. The highest reliability coefficients of these studies were generally in the .30's and .40's. These studies include Heron's (4) studies of the inclined plane problem-box and the Watson circular maze with rats in 1922, Heron's (5) study of the stylus maze in 1924, Hunter's (7) study of two stylus mazes with human subjects and of three mazes of different complexity with rats in 1922, Hunter and Randolph's (9) study of the reliability of several stylus mazes and of the intercorrelation between records on a maze, a straightaway and a problem box with rats in 1924, Hunter and Randolph's (10) study of the reliability of a very simple maze with the goat in 1926, Liggett's (16) study of two simple mazes with the chick in 1926, and studies by Tolman (26) in 1924 and Tolman and Davis (27) in 1924 of several relatively simple mazes with rats.

A comparison of these earlier experiments with more recent ones yielding higher coefficients seems to indicate that the features of the early experiments responsible for the low reliability coefficients were: (1) the fact that the mazes were too simple and easy, (2) the lack, in most cases, of preliminary training to accustom the animals to the apparatus and handling and to develop stronger motivation, (3) poor control of motivation, (4) the use, in some cases, of mazes with alleys of such unequal complexity that chance blunderings into certain alleys offered much greater hin-

drances to learning than blundering into others, (5) the lack of means of preventing retracing, and, (6) utilization of too few trials to furnish the data correlated. In some experiments, for instance, the correlated scores were errors on different single trials; in other cases, groups of only three trials were correlated. Just as with psychological tests, the reliability coefficients from such small units are generally lower than from units providing more chances for success or failure.

Within the last three or four years a number of studies of maze reliability have been reported with much more positive results. The first of these was the exploratory experiment of Tolman and Nyswander (28) in which, of the variety of patterns used, the multiple-T maze since used by Stone and Nyswander (25), Heron (6), and by the present experimenter, yielded definitely the highest reliabilities. [It is interesting to note that in 1922 Hunter (7) had found a simple T maze the most reliable of the three mazes he used.] Stone and Nyswander (25) have given this simple T maze a much more extensive test in connection with their study of the influence of age on learning. Reliability coefficients, figured on eight groups of about 25 rats each, were of values of about .80 to .90. Some doubt, however, is thrown upon these figures by a recent study by Heron (6). Duplicating Stone's technique, in general, Heron found quite as high reliability coefficients from correlation of different parts of the record of the original learning or of a relearning; but, on retesting his groups after intervals of 221 and 175 days (with two groups of 36 and 54 rats, respectively),

he found that the correlations between scores in the original learning and scores in relearning were .376±.07 and .326±.06. Within that period of time, of course, accidental factors may have affected differently the learning ability of different rats, so that perhaps the correlations are legitimately lower, as one might say. But Heron's experiment seriously raises the question as to whether Stone and Nyswander's reliability coefficients may not have been so high partly through the existence of certain factors essentially of the nature of systematic errors which tend to produce the high internal-consistency correlations. For example, systematic differences between the animals in feeding or emotional conditioning to handling and to the apparatus might persist through a 30-day period of training and so produce a consistency of performance as between the different parts of the original learning or of the second learning; but they might not persist through as long a rest interval as in Heron's experiment.

The experiment by Tryon (30, *33a*) illustrates rather clearly the possibility of this source of error. On the first of the two mazes on which his rats were trained, the procedure was adopted of feeding all animals of the same sex the same quantity of food, regardless of the fact that the age range was from 3 to 8 months and that the food requirements of the different rats must have varied widely. It would seem that, inevitably, the consequence of this feeding program would tend to be a consistently different rate of learning for different rats *based on the differences in motivation.*

It should be added with regard to Tryon's experi-

ment, however, that, with a group of 107 rats, training was *not only* given on his first (mechanical) maze, but also on a second (hand-operated) one; that on this second maze the feeding program was quite different; and that, nevertheless, high correlations were found between the scores on the two mazes. On the second maze, Tryon says, ". . . the procedure of letting each animal eat until he first turned away from his food pan was adopted." In spite of this change of feeding program, when errors from groups of three trials on the first maze were correlated with errors on groups of three trials on the second (with 20 trials given on each maze and the first trial and the twentieth dropped), the resulting raw correlation coefficients ranged from .318 to .772, with the median coefficient .608. When the successive groups of three trials on the first maze were correlated with Trials 2-19 on the second, the correlation coefficients were .470, .639, .704, .749, .773, and .793. The correlations between the successive groups of three trials on the second maze with Trials 2-19 on the first maze were .758, .709, .712, .600, .602, and .585. The change of feeding technique, and the 7 or 8 days' period between the two periods of trials quite materially reduce the danger that these correlations have been raised by systematic errors of feeding. Nevertheless, one still cannot help noticing that there is a consistent tendency for the correlations to be higher the fewer the days separating the correlated trials. This seems rather definite proof that there must still remain a tendency for adjoining groups of trials to be affected by common factors other than maze learning ability, be-

cause it certainly would not seem that the true ability of these rats had changed so in the course of the 45 or 46 days involved in these correlations.[1] [This is supported by Lashley's (14) findings which show that an increase in the time interval between tests brings with it an increase in the *diversity* of errors made and, consequently, a decrease in the errors common to the tests.]

The other recent experiments which have demonstrated rather satisfactorily high reliability coefficients are the studies by Husband (11), Liggett (18), and Yoshioka (35, 36, 37). Each of these experiments used a different style of maze, and any one of the patterns might well have been selected for the further investigation of this study. Tryon's mazes, of all the group, seem the ones that probably offer the best features; but his pattern was not taken because of the fact that the

[1]Since the above was written, several articles by Tryon have been published which deal with some of these criticisms. Thus, in "Individual differences in maze ability: II The determination of individual differences by age, weight, sex and pigmentation" (33*a*), he has shown that none of the variables discussed in the article was significantly related to maze performance. Several correlations of error scores and weight were figured for the 88 male rats used. The correlation between weight at the start of training and scores on the first maze learned was —.09, that between scores on the second maze learned and weight at the conclusion of training was —.11. I still remain skeptical, however—partly because of the data from the present study that show that reliability coefficients can be raised by differences of feeding. Moreover, the description provided in these articles of the retracing doors makes the problem of differences in emotional reaction appear as possibly quite a significant factor. The doors were treads inclined at an angle of 45° from the floor, each suspended by a rubber band in such a fashion that the door would sink to the floor as the rat walked on it and be jerked back up as soon as the rat had passed over it. Such an arrangement would seem to me too liable to produce a differentiating emotional conditioning in various members of the group.

present work was started before information on his study was secured. The multiple-T maze was selected because of the reason that its relatively high reliability had been demonstrated with larger groups than were used with the other mazes.

In summing up the work of the past, it clearly seems *that with different patterns and different experimental* procedures vastly different results as regards reliability can be secured. Some light has been thrown on the question of what types of maze pattern and experimental procedure produce different reliabilities, but the more exact determination of the influence on reliability of different possible variations still offers a field for fruitful work.

III

THE LOGICAL AND STATISTICAL PROBLEMS OF RELIABILITY AND VALIDITY AS RELATED TO MAZE EXPERIMENTS

A. The Formulae for the Reliability of Differences between Group Averages

A considerable part of the discussion about the problems of the reliability of maze experiments has been concerned with the formula for the standard error of a mean:

$$\sigma_M = \frac{\sigma_{dis.}}{\sqrt{N}}$$

and the formula for the standard error of a difference:

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2} - 2\, r_{12}\, \sigma_{M_1}\, \sigma_{M_2}} \qquad [I]$$

(or, where the data entering into the determination of the two means cannot be correlated:

$$\sigma_{M_1 - M_2} = \sqrt{\sigma^2_{M_1} + \sigma^2_{M_2}} \qquad (12, \text{p. } 182) \ [II]$$

Apparently, Formula I is sometimes regarded as being a rather questionable statistical refinement of Formula II, as though Formula II were really the legitimate formula, and Formula I somewhat on a par with representing the index of a reliability ($\sqrt{r}$) as indicating directly the reliability of a test. The fallacy of this view is well disclosed in Walker's article (34). To her algebraic treatment of the problem, however, the suggestion might well be added that the general reason why Formula I (with data which can be cor-

related) is no more radical than is Formula II for *its* appropriate data (data which cannot be correlated) is this: In the case of data which can be correlated, errors of sampling cannot enter in to affect the difference between means, as they can in the second case. Thus, suppose that our task is to discover the effect on error scores of doubling the feeding time at the end of the tenth trial. If we use the same group throughout, as naturally would be done, there can be *no* errors of sampling to affect the difference in means of scores of the trials preceding and following the tenth trial. If, however, we run one group for 10 trials and compare its record with that of a second group subsequent to the tenth trial, the difference between the means can be affected not only by errors of measurement, but also by errors of sampling. Hence, in this latter case, we not only have to use Formula II, but we find it the correct formula to use.

Regarding these formulae, the discussion has centered upon three main issues: (1) When the reliability of a particular instrument is low, are these formulae sufficiently conservative? (2) Is the situation quite the opposite, as Tryon has claimed, and is a new formula needed to correct for too great a conservativeness with measurements from unreliable instruments? (3) Do maze data sufficiently conform to normality of distribution to cause means and $\sigma_{M_1 - M_2}$'s to be the best statistical measures by which to measure the differences between groups?

The discussion of the first issue mainly concerns the controversy between Carr (1) and Hunter (8) and

several of his students. The problem has been given a very satisfactory treatment, however, by Tolman and Nyswander (28), and need not be gone through again. These authors point out that the problem is simply one of practicability—that dependable quantitative results can be secured with instruments of very great unreliability, but that either the difference produced by the experimental factor will have to be so great, or the groups will have to be so large, before the ratio $\frac{D}{\sigma_{diff.}}$ reaches satisfactory proportions, that the discovery of more reliable methods is an urgent problem. Of course, this formula is not the only safeguard needed, even where the data have fairly normal distribution. The existence of any bias in the selection of the samples, for instance, might produce groups yielding statistically different means. The point is, however, that not only will the $\frac{D}{\sigma_{diff.}}$ formula fail to detect such errors, but knowing the reliability coefficients of the instrument will not help either. Such sources of error can be guarded against only by the adequate control of non-statistical phases of the experiment.

The second issue is the question of the merit of Tryon's proposed new formula (for the $\sigma_{M_1 - M_2}$), which has won favorable comment from at least several authors. The consideration of this issue has been rendered somewhat difficult by the fact that, in the course of his three articles (29, 31, 32), Tryon has shifted his position twice, without indicating at all clearly his renunciation of his earlier different positions. He at first announced that he was demonstrating

that, when a certain *difference* was found between two means, that *difference* was to be taken as more significant the greater the unreliability of the measuring instruments involved. This was then renounced for the proposition which his statistical argument all along had been designed to prove, namely, that when a certain *ratio* was found between a difference and the sigma of that difference, that *ratio* was to be taken as more significant the more unreliable the measuring instrument. His third article seems to foreshadow a re-acceptance of the ordinary $\sigma_{M_1 - M_2}$ formula.[2] It is the second of his propositions that calls for discussion here.

In criticizing Tryon's arguments, it is helpful to get a picture of how the ordinary $\frac{D}{\sigma_{M_1} - \sigma_{M_2}}$ formula operates and of its relationship to the reliability of the test. First, it is to be noted that, with greater and greater unreliability, the difference found between two groups may vary more and more from the true difference than would be the case with a perfectly reliable instrument. Thus, if the true difference is 4 points, with a measuring instrument of a reliability of .80 one might occasionally find differences of 2 or 6 points

---

[2]In a still more recent article, Tryon writes thus: "Applied to maze measures, my notion was, briefly, that when a difference between mean maze scores of two groups who differ in some systematic way *has been found* to be statistically reliable by gauging it in terms of the orthodox $P.E._{diff.}$ formula, the reliability of this difference is unaffected by knowledge of the reliability coefficient within each group" (33*a*, p. 156).

It is interesting to know that this is what Tryon *meant* by his earlier articles; when one reads the articles in question one tends rather to conclude that Tryon has rather fundamentally changed his opinions.

rather than 4; and with a measuring instrument with a reliability of .50, differences of —1 and 9. This can be readily realized if two groups whose true means are absolutely equal are considered. With a perfectly reliable test, no difference would be found between the means, but with tests of greater and greater unreliability there would tend to occur either positive or negative differences of greater and greater size. There is, accordingly, no safeguard in the numerator of the $\frac{D}{\sigma_{M_1 - M_2}}$ formula to guard one from accepting as true those differences that tend to be farther and farther from the true values as the measuring device becomes more and more unreliable. In the denominator of the formula, however, there is such a safeguard—namely, that, as the reliability decreases, not only does the mean tend to deviate farther and farther (one direction or the other) from the true position, but also the sigma of the distribution tends to be increased. It is generally held that this size of the $\sigma_{dis.}$ tends to increase with unreliability at a rate which compensates for the tendency of the mean (because of unreliability in measurement) to assume positions away from the true mean. The essence of Tryon's suggestion was, however, that instead of using the $\sigma_{dis.}$ directly, in the $\sigma_{M_1 - M_2}$ formula, a formula derived by Kelley for estimating the probable true sigma

$$\sigma_{True} = \sigma_{dis.} \sqrt{r}$$

should be used, thus giving the formula:

$$\sigma_{M_1 - M_2} = \sqrt{\left(\frac{\sigma_1 \sqrt{r_{11}}}{\sqrt{N_1}}\right)^2 + \left(\frac{\sigma_2 \sqrt{r_{2II}}}{\sqrt{N_2}}\right)^2}$$

The values achieved by this formula will be the same no matter what the reliability of the measuring instrument, Tryon said (31, p. 4), inasmuch as the formula " . . . . contains the true sigmas of the given groups which are necessarily constants, otherwise, they would not be 'true.' " Such being the case, one would accept a difference between means of 4 points, let us say, as being just as reliable when secured with tests with a reliability of .20 as when secured with tests with a reliability of 1.00. But, as we have seen, the more unreliable the test the greater is the probability that the found difference between the means may be either greater or less than that which actually should be indicated.

The difficulty in Tryon's reasoning was indicated when he specified his reasons for preferring the formula for the true sigma rather than the ordinary $\sigma_{dis.}$ for use in the $\sigma_{M_1 - M_2}$ formula. He said:

> "The *true sigma* is an index of the dispersion in a group when each individual's score contains absolutely no error of measurement. Obviously, this is the sigma which we are fundamentally interested in and is the one to be used in comparing populations with each other. . . " (31, p. 2).

The defect in this statement is that, when we are trying to determine the significance of a mean or of a difference between means, it is not the true sigma in which we are fundamentally interested, but, on the contrary, the actual or found sigma of distribution, since it is only this latter sigma which has any influence in the $\dfrac{D}{\sigma_{M_1 - M_2}}$ formula in warning against possible displace-

ment of the means from their correct positions because of the unreliability of the test.

The third issue which has been raised relative to the formula for the critical ratio, and the issue most deserving of further attention, is the question of the allowances to be made for the general tendency of maze data to deviate markedly from normality of distribution.

Paterson in 1917 called attention to the fact that the maze studies he reviewed contained some markedly skewed data, and suggested that, in view of this fact, the median might have certain advantages over the mean as a measure of the central tendency of the group. This suggestion has won little application. Probably a reason for this is the unwarranted belief on the part of so many psychologists that the mean is in every case superior to the median as a statistical tool except where ease of calculation is the prime desideratum. With maze data, as well as with quite a quantity of other psychological data, however, the median is really the statistically preferable tool on the grounds of smallness of standard error. Liggett in 1928 indicated the fundamental reason for the unsatisfactoriness of the mean relative to maze data:

> "When results like this are secured, the curve must always be skewed, for it is impossible to get low scores that will counterbalance the high ones, due to the nearness of the physiological limit at the lower end of the scale. A greater number of cases will not change this, for one may get more high scores, but cannot get scores small enough to cancel the high ones. . . . .
>
> "Quite possibly the maze is a reliable method of measuring learning, but on account of the peculiar distri-

> bution of learning records, the present statistical methods of treating maze data do not give significant group differences. If reliable conclusions can be drawn from group averages in the maze, it will be after more knowledge is gained of distributions, and of methods of treating skewed curve data.
>
> "Granting that maze records do not distribute upon the normal curve, the arithmetic mean is not an accurate measure of central tendency. . . . " (17, pp. 54-55).

Again, in Yule we find:

> "For a normal curve the standard error of the mean is to the standard error of the median as 100 to 125, and in general the standard errors of the two stand in a somewhat similar ratio for a distribution not differing largely from the normal form. . . . The mean is *in general* less affected than the median by errors of sampling. At the same time, we also indicated the exceptional cases in which the median might be the more stable—cases in which the mean might, for example, be affected considerably by small groups of widely outlying observations, or in which the frequency-distribution assumed a form resembling fig. 53 (i.e.—flat-topped curves), but even more exaggerated as regards the height of the central 'peak' and the relative length of the 'tails.' Such distributions . . . . might be expected to characterize some forms of experimental error. . . . .
>
> "*Further, in some experimental cases it is conceivable that the median may be less affected by definite experimental errors, the average of which does not tend to be zero, than is the mean,*—this is, of course, a point quite distinct from that of errors of sampling " (38, pp. 344-345). (Italics mine.)

The condition mentioned in the last sentence of this quotation would seem to be the condition with maze data, as Liggett has pointed out. One cannot expect that experimental errors on such learning problems as this will cancel in their effects. The solution of the

matter, however, is probably not as difficult as the quotation from Liggett suggests. The median can be taken as a fairly satisfactory measure of central tendency even under these conditions, for it is reasonable to assume that the majority of the animals will not have been affected by serious experimental errors and the value of the median will be determined by this large majority of unaffected animals.

B. Problems of Validity

Only limited experimental data exist on the question of validity of different methods of measuring learning ability in animals, in spite of the fact that such information is basic for the accurate interpretation of experimental results.[8] It may be true that the ability to eliminate errors or excess time on a maze is indicative of learning ability in general, but this cannot be assumed without proof. The very favorable maze performance of rats in comparison with the maze performance of human subjects, coupled with the very obvious differences between the two on other learning problems, suggest that maze-learning ability is certainly not an index of all types of learning ability. Most experimenters on animal behavior, however, instead of securing experimental evidence of the validity of their measures, have been content to assume that the learning scores on a maze, or light-discrimination problem, or problem box, are indicative of "learning ability" in

[8] A recent article by Tryon (33*b*), "Studies in individual differences in maze ability. III. The community of function between two maze abilities," gives a good summary of the experimental material available to date on this question.

general. To what extent this is true cannot be estimated with any assurance until one group of animals has been tested not only on one such particular problem, but also on a considerable variety of other learning problems, and until the correlations have been determined between the scores on that particular problem and the scores on the other problems. Each of these correlations would be a validity coefficient of scores on either problem as a means of indicating the probable scores that the animals would make on the other problem, and the series of correlations between scores on one particular problem and scores on a number of other problems (varied enough to sample all the fields of learning) would be the means of estimating the validity of that particular problem as a means of estimating "learning ability" in general.

It is obvious from what is said that a given measuring instrument may have any number of validities, because, after all, a test is not simply valid in some abstract sense—tests have validities only relative to such and such other things. Thus, a given maze might be found to have one validity as an index of the ability of rats to learn such and such other varieties of mazes; still another validity as an index of ability to solve problem boxes; and still a third validity as an index of the ability of the animals to make fine visual discriminations.

The consideration of this concept of validity coefficients indicates clearly that only a vague and more or less arbitrary line separates "validity coefficients" from "reliability coefficients." For, while reliability coef-

ficients may be defined theoretically as the correlations between two measures which sample the same functions in the same proportions in each case, practically speaking, reliability coefficients are ordinarily calculated from measures which only approximate this ideal, and they really sample slightly different ranges of functions. However, as validity coefficients become more and more restricted in their scope of reference, they also approach this same status. So when, as in the present study, certain of the correlations are of scores on two different mazes, it is a more or less arbitrary choice as to whether one calls the coefficients validity coefficients or reliability coefficients. Strictly speaking, they are validity coefficients, but, practically speaking, the two measures measure ranges of functions that are so nearly identical that they can actually be used as reliability coefficients.

C. Problems of Reliability Coefficients

In evaluating the different possible methods of figuring reliability coefficients, we will be aided if we keep in mind the concept of reliability as being the measure of the extent to which chance factors have been excluded from obscuring the measurement of some more or less fundamental characteristic of the members of a group. Measurements of reliability in this sense could be had directly, if it were possible to secure two samples of performance which would satisfy the requirement of being absolutely independent measures of the same thing. In studies of behavior, it is rare that reliability coefficients can be taken as altogether accurate indexes of the reliability of the measur-

ing instruments involved, either because they are not independent measures or else because they are not entirely measures of the same thing.

A few illustrations have already been given of the ways in which reliability coefficients may be defective, but the entire matter may be summarized here now. In the first place, it is possible to cite cases where the reliability coefficients are higher than the reliability of the measuring instrument really justifies. For example, where the animals of a group are not equally motivated, they tend to be consistently different because of this difference in motivation. Moreover, from the standpoint of most maze experiments, such differences resulting from poor control of feeding are the result of essentially irrelevant factors, and the correlation of odd and even trials, for example, is not a correlation of two independent measures of the same thing, but of two measures between which certain systematic errors exist which raise the correlation. On the other hand, the lack of independence between the two measures may operate to reduce the correlations. In the present experiment, negative correlations were found with certain groups between errors on the first trial and errors on the remaining trials. The most reasonable interpretation of this is that the scores on the first trial were largely determined by chance, but the subsequent performance tended to be improved if a rat had, by chance, made an unusually large error score on the first run. In this case, the reliability coefficient will be lowered rather than raised as a result of the lack of independence of the measures.

It is highly probable that correlations of odd versus even trials or of such parts of learning as Trials 1-10 versus 11-20 on the same maze have been influenced by such carry-over as above described. For this reason, the test-retest method of calculating reliability coefficients is of particular interest. Particularly where the animals are returned to normal feeding and where a considerable period of time is allowed to intervene between test and retest, there is relatively little possibility that systematic experimental errors will last over, or that transfer effects will give rise to spurious correlations. This is particularly true if the retest is given on a maze with a different pattern.

Where the test and retest are separated by too long an interval of time, the resulting reliability coefficients may be defective in a different way, namely, that while they are independent measures, the same thing does not exist to be measured in the two cases. In other words, with intervals as long as those used by Heron (6), the maze ability may change and the reliability of the maze as a measure of the learning ability existing in either period may really be higher than the test-retest coefficient would indicate. The conclusion would seem to be, then, that with maze studies it is perhaps impossible to get reliability coefficients which are not more or less biased in one direction or another. To estimate reliability, therefore, the procedure that seems necessary is to calculate reliability coefficients in a number of ways, using those methods of calculation which most approximate the ideal of having independent measures of the same thing, and then estimating reliability from the

total collection of reliability coefficients, trying to make due allowance for any systematic errors which may have affected the correlations.

Let us turn, then, to a critical evaluation of the different methods of calculating reliability coefficients which have been suggested during the history of the problem: Stone and Nyswander have used the following methods. (*a*) correlation of scores on odd versus even blinds; (*b*) correlation of scores on the first versus the second half of the maze; (*c*) correlation of scores on odd versus even trials, and (*d*) correlation of scores on different groups of trials. All of these methods may be characterized as measures of the internal consistency of maze scores. Methods *c* and *d* have been used in the present experiment, but Methods *a* and *b* are rejected because these correlations are probably too seriously affected by systematic errors. An error made by chance on one blind, for instance, may tend to confuse the rat for some blinds following. There is not so great a probability, however, that scores on successive days or successive groups of trials would be quite as lacking in independence.

The third method used in calculating reliability coefficients in the present experiment is the test-retest method. In addition to the merits suggested above, in the previous discussion of this method, an additional merit comes from the fact that one is interested in knowing the reliability of the entire series of maze trials. With the test-retest correlation, one can take the raw correlation coefficient as indicative of the reliability coefficient of either test separately when the

assumption seems warranted that the test and retest have both had about the same reliability. On the other hand, if one is trying to calculate the reliability coefficient of an entire series of trials from the various internal-consistency correlations, one must use the very dubious procedure of calculating the reliability coefficient of the series with the use of the Brown-Spearman formula. It has been empirically demonstrated, of course, that this formula predicts accurately the effects of lengthening certain types of educational tests within certain limits of increase of length. However, this formula is peculiarly dependent on the assumption that the added material is of equal difficulty, and independent of the material to which it is added. And as has been indicated before, it seems certain that the various parts of one training series do *not* have this independence.

Still another method has been suggested and used by Lashley (1929). Lashley's procedure and the merits he claims for it, are indicated in the following:

> "To avoid overlapping of the data and consequent spurious correlation, the average time per trial and the average errors per trial for the first 10 trials of learning were correlated with the total number of trials required to reach the citerion of learning. The measures correlated are thus mutually exclusive" (15, pp. 20-21).

Whether these merits actually hold for this method, however, depends on whether any appreciable number of animals meet the norm of mastery before the 10 trials are completed. If such is the case, there actually may be some element of spurious correlation due to the fact that errors in the latter part of the learning curves

of such animals will tend to raise scores on both types of measurements. This defect, however, can be avoided by correlating errors from such a limited group of trials that only a few animals, at the most, will have met the norm of learning in that period. Even with this precaution, however, one fact that still remains is that the correlated measures are from the same training series and are therefore presumably subject, to some extent, to the same defects as the other measures of the internal consistency of a maze experiment. Still another objection to this method of figuring reliability, however, is that it depends upon the correlation of measurements of two different types, and that unreliability of either type of measurement will make the reliability coefficients from the combination rather lower than the reliability of the better of the two.

Still another method of correlating scores, to determine the extent to which chance factors were influencing maze scores, used by Hunter (7) and Heron (4), is the procedure of correlating Vincent scores rather than raw scores. Just how the resulting reliability coefficients are to be interpreted, neither Hunter nor Heron has indicated, and the fact that Vincent curves are of the highest value for certain other problems (namely, where one wishes a picture of the form of the learning curve for a group of subjects) does not insure that they are useful here. At least one serious objection to their use can be pointed out. This objection is that converting raw scores into Vincent scores spuriously raises the correlation by making the scores in any one tenth dependent on the total number of trials required by

that subject to attain the norm of mastery. Any chance factor which increased the number of trials required by the rat to reach the norm of mastery would affect the scores through all the tenths of learning. Hence, correlations of Vincent scores would be affected by some unknown and more or less unmeasurable element making for spuriousness in correlation. On the other hand, with correlations of odd versus even trials, while it is true that chance errors may affect the two series of scores correlated, in those cases where the effect of the chance factor endures for more than one trial, there is nothing in the mathematical treatment of the data which would tend to make for spurious correlations, as is the case with Vincent scores.

In addition to the precautions suggested above, a most important consideration in the comparison of reliability coefficients derived with different groups is the consideration of the influence of different ranges of talent in the different groups. This is important because of the fact that, with the same maze pattern and maze procedure, the size of the reliability coefficients secured will be to a major extent a function of the range of true ability in the groups. The more heterogeneous the group, all other things being equal, the higher will be the reliability coefficients.

As Kelley and Shen state:

> "As the range of a distribution increases, ... errors of estimate remain relatively constant, and consequently r will increase with the standard deviation. Assuming, then, that errors of estimate are equal for different ranges, we have the following relation between the dispersion of the distributions and the magnitude of the correlation coefficient:

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1-R}}{\sqrt{1-r}}$$

where r is the correlation coefficient derived from distribution with a standard deviation of σ, and R from distributions with a standard deviation of Σ. Since the range of data is more or less arbitrary and accidental in any given study, it should always be taken into account in interpreting the correlation coefficient derived" (21, p. 849).

The clearest illustration of this tendency that is available from maze data is found in the comparison by Lashley (15, pp. 20-22) of the reliability coefficients of his maze when figured, first, from the records of normal animals, and then from the records of a group with brain lesions. The reliability coefficients found by means of correlating total trials to learn with time and errors in the first 10 trials are presented in Table 1. In correlating scores for time, errors, and trials on two mazes, Lashley found the correlations for the two groups as shown in Table 2.

Of course, it must be admitted that between his normal group and his operated group there was a dif-

TABLE 1

RELIABILITY COEFFICIENTS FROM LASHLEY'S EXPERIMENTS AS FIGURED BY CORRELATION OF TRIALS TO LEARN WITH TIME AND ERRORS, RESPECTIVELY, IN THE FIRST TEN TRIALS

Scores correlated	N in group	Reliability coefficients Time	Errors
Learning scores of normals	59	.09±.09	.06±.09
Learning scores of operated rats	37	.18±.11	.62±.08
Retention scores of operated rats	59	.76±.04	.79±.04

TABLE 2

Correlations Found by Lashley between Scores on Two Different Mazes with Normal and with Operated Rats

Scores correlated	Normal group	Operated group
Time	—.09±.18	.55±.09
Errors	—.36±.16	.67±.07
Trials	—.38±.15	.68±.07

ference in range of ability such as might never be found between two groups of normal animals, nevertheless the probability of differences in range of ability in different groups is sufficiently serious so that consideration of this factor is important. In Tryon's (30), experiment, for instance, because of the fact that Tryon was primarily interested in studying the inheritance of maze-learning ability, the rats were selected so as to yield as heterogeneous groups as possible. It is reasonable to expect, therefore, that the reliability coefficients which he secured are higher than would be secured by another experimenter using the same maze and exactly the same procedure, but groups of rats as homogeneous as those used in Stone and Nyswander's experiment or in the present experiment.

Just how much higher the reliability coefficients are as a result of unusual heterogeneity of subjects in maze experiments cannot be estimated, however. The procedure suggested by Kelley and Shen for estimating the influence of range—the procedure of using the formula:

$$\frac{\sigma}{\Sigma} = \frac{\sqrt{1-R}}{\sqrt{1-r}}$$

—is not applicable to maze data because of the fact (as the present experiment has demonstrated) that the assumption mentioned in the above quotation (6.000) is not satisfied with maze data. Errors of estimate *do not* remain relatively constant under the different conditions under which different experiments are run. The errors of estimate, as will be shown later, vary with the same group in different portions of learning, and differ between groups for the same set of trials whenever any such factors as differences in feeding exist between the two groups. With test scores it is mainly true that only differences in range of ability sampled and errors of measurement will cause differences in the sigmas of different distributions. With maze scores, however, the sigma of scores for a given portion of learning will be determined not only by the degree of heterogeneity of the group and by the unreliability of the measuring instrument, but also by whatever factors tend to produce a high or low mean score for the group for that portion of learning. Hence, if we applied the formula suggested by Kelley and Shen, the groups with the lowest learning curves would be indicated as having the greatest reliability simply because of the fact that (due to stronger motivation or better preliminary preparation, perhaps) their scores are more closely grouped and the subjects appear to be a very homogeneous group. To take an extreme case, if learning has been carried on long, most of the group may be making zero scores, and the distribution of scores will be extremely small; but this is obviously not dependent upon great homogeneity of the group.

The difficulty, then, is to get the variability coming from such sources as motivation separated from the variability coming from differences in ability and from variability due to unreliability of measurement. Some sort of direct measure is needed of the amount of variability due to differences in ability, for the portion of the variability due to such differences in range of ability should be taken out; but the variability determined by other factors should be left in as part of the experimental data.

There is a formula suggested by Garrett (3, pp. 276-277)

$$\frac{\sigma \sqrt{1-r}}{M}$$

which might be used for comparing the reliabilities of different maze groups, if it were true that whatever factors affect the sigma of scores of a group would affect the mean of the scores in the same proportion. (It is to be noted that this formula of Garrett's is merely a cross between the formula for the standard error of a score:

$$\sigma_{score} = \sigma \sqrt{1-r}$$

and the coefficient of variation

$$V = \frac{100\,\sigma}{M}\Big).$$

Wherever this condition holds, this formula makes possible the comparison of the reliabilities from different groups in the same manner as the ordinary formula for the standard error of a score permits the comparison of

reliabilities from different groups when the errors of estimate with the different groups are known to be equal. However, with maze data there does not exist this close and direct relation between the sigma of score for for any portion of learning and the mean of the scores for that same portion—the factors that affect the mean affect the sigma in a different degree in different cases. Thus, when the coefficients of variation for the group of 107 rats of Tryon's experiment are calculated, the coefficients for the successive groups of 3 trials on the first maze are 27, 63, 78, 90, 95, and 105; and on the second maze, 40, 57, 80, 77, 91, and 88. Inasmuch as the range of true ability was that of the same group throughout, and inasmuch as the reliabilities of these different periods was not as different, by any means, as these figures are, one is forced to conclude that the factors that affect the means with maze groups do not affect the sigmas to the same degree, and that, consequently, even this formula suggested by Garrett can hardly be used for the comparison of the reliabilities found with maze groups having different ranges of ability. It may be that, where differences in range are extreme, comparisons may be made slightly better through the use of this formula, but, with the differences in range more commonly existing, the use of this formula would probably lead to greater errors than it would cure.

Hence, the only way to safeguard against the possibility that reliability coefficients from different maze groups have been made incomparable by differences in ranges of ability is to use groups that are known to have approximately the same degree of homogeneity. Using

stock that is highly inbred as the source of all the groups gives an approximation to this ideal; but the only really satisfactory procedure seems to be the split-litter technique, the importance of which has been experimentally demonstrated in so beautiful a fashion by Corey (2) in 1930. The benefits made possible by the use of this technique, as Corey points out, are not only the rough equating of groups with respect to heredity, but also the equating of the groups with respect to seasonal variations in temperature and humidity and all the other changes which more or less inevitably slip in, despite the experimenter's attempt to keep conditions absolutely constant. Strong corroboration of Corey's points is to be found in the data of the present experiment. So important is this consideration that I would venture to say, with some assurance, that there is no means today whereby an experimenter can claim that he has with reasonable certainty found that a certain maze pattern or maze procedure yields greater reliability than some other pattern or method used by others except by using the procedure of the split-litter technique to compare the reliabilities of patterns or methods in question. This, of course, would not hold where experimenters had used approximately similar conditions and stock and yet had found extreme differences in the reliabilities, but would hold quite definitely, I believe, when one seeks to estimate whether the procedures of Tryon, Husband, Liggett, Stone and Nyswander, Yoshioka, or the present experiment yield the more reliable results.

IV

EXPERIMENTAL RESULTS

A. Introduction

The first object of this experiment was to determine whether, with the multiple-T maze, the highest reliabilities could be secured with maximum, moderate, or minimum prevention of retracings. The second object was to determine whether a program of strong restriction or a program of but moderate restriction of feeding would yield the higher reliability. The third object of the experiment was to throw additional light on the problem of the interpretation of reliability coefficients from maze experiments. To this end various treatments have been worked out of the scores of different groups on a single maze, and also three of the groups have been given retests after a fairly long rest interval following their original training, and the test-retest correlations calculated.

The maze pattern used was essentially that of Stone and Nyswander (27) except that, due to the fact that their specifications of dimensions were not clear, the present maze was constructed with each unit 4" longer than theirs. (The correct dimensions of their maze have been secured from Stone directly.) One trial was given each day, and Stone and Nyswander's methods were duplicated as closely as possible, with the major exception that the preliminary training was given on a short straightaway rather than on the simple one-tread problem box which they used, and for a 10-day period rather than their 5-day period.

B. Apparatus and Method

The apparatus and procedure of the present experiment will be described in some detail because our knowledge is still so scanty regarding the factors influencing reliability that when we have found a certain reliability from one particular experiment we cannot say with certainty which factors have been responsible.

1. *The Straightaway Used for Preliminary Training.* For preliminary training, a small straightaway was used. The starting box and the food box of this were similar to those of the regular maze, and the alley was the same in construction except that there were no turns or blinds. For use in the first three days of the preliminary training with certain of the groups, the length of this straightaway was 26″ from the exit of the starting box to the entrance of the food box. For the remaining trials, with these same groups, and for all of the trials with all the other groups, the straightaway had a length of 6′.

2. *Construction of the Maze.* The pattern of Maze I is shown in Figure 1. Maze II was a mirror image of Maze I. It was constructed by turning Maze I upside down and attaching the wire covering to the new top of the maze. The maze was constructed of ½″ white pine boards, and was painted a flat black on the inside. The alleys were 4″ wide and 4″ deep, inside dimensions. The ceiling of the maze was made of ½″ hardware cloth nailed to the top of the walls of the maze. The maze had no floor of its own, but rested on the floor of the room. This floor was covered with varnished brown battleship linoleum which was cleaned

after each day's work by mopping it up with a wet cloth. To facilitate this cleaning, the maze was attached to ropes in such a way that it could be lifted as a unit to a height of about 3′ from the floor.

Retracing doors were hung at the points indicated in

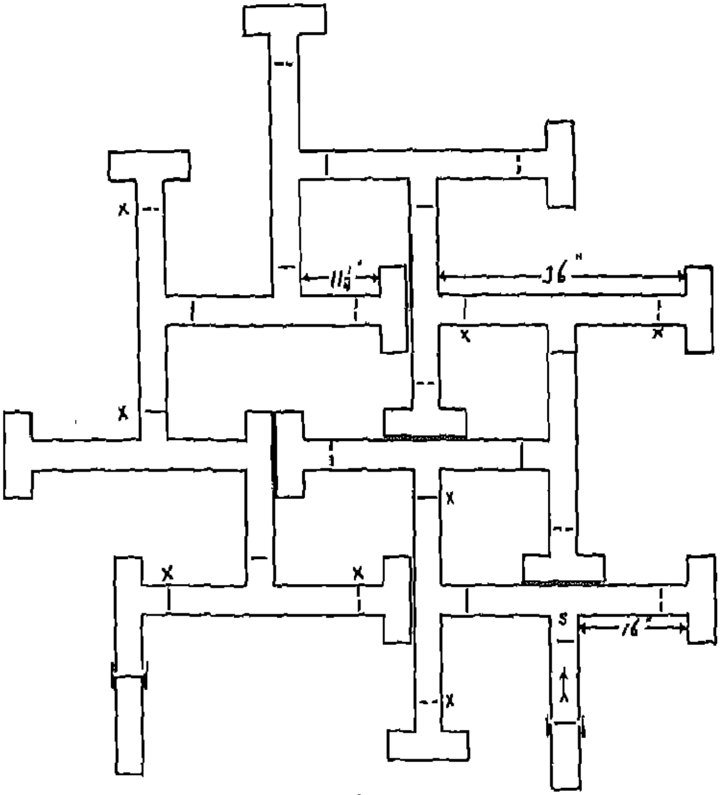

FIGURE 1

DIAGRAM OF MAZE I

The doors used to prevent retracing are indicated by solid lines across the path, the pseudo-retracing doors are indicated by broken lines across the path. The doors used with the 4-door groups are indicated by *x* marks. With the 1-door groups only the door in the last alley was used.

Figure 1 by solid lines across the maze path. These doors were constructed of hardware cloth stretched on wire frames and were so hung from the top of the maze

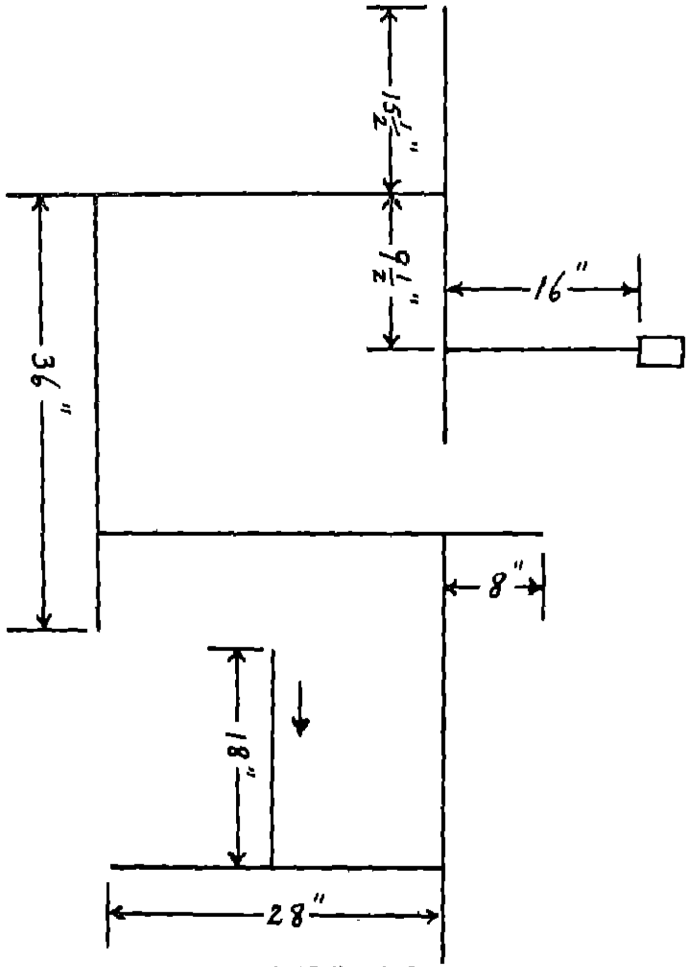

FIGURE 2
DIAGRAM OF THE ELEVATED MAZE
Width of path, 1⅛"

that they could be lowered to prevent retracing. These doors were operated from outside the room. At the other ends of the alleys in each unit dummy doors were hung so that the rats could not learn to make their turns on the basis of visual cues derived from the doors. The rats were released from the starting box by a door sliding in guillotine fashion. To avoid all possible noise during the run, this door was left open until the rats had reached the food box.

The third maze used was an elevated maze (Figure 2). The width of the path was 1⅜" and the height of the path from the floor was 21". No starting box was used with this maze. The rats were merely placed by hand at the starting point. This elevated maze was not located in the sound-proof room (to be described below), but as it was placed in a relatively secluded room of the laboratory, and, as all of the trials on it were given in the evening, there was relatively little noise attendant on trials with it. In running the rats on this maze, the experimenter stood at a small, high-topped table about 4' from the maze. No screen was used to conceal the experimenter from view. Each rat was carried into the room by hand, a distance of about 25', and was returned to its cage before the next rat was secured.

3. *The Room Used to Control Sound During Runs.* All of the trials on the straightaway in the preliminary training and all of the trials on Mazes I and II were given in a room which was approximately sound-proof. The main details of construction of this room are shown by the diagram, in Figure 3, of a corner of this room.

For each trial, the rat was carried into this room and left there alone until the completion of the run. Errors were recorded by the experimenter from outside the room.

The walls of this room were constructed of Celotex,

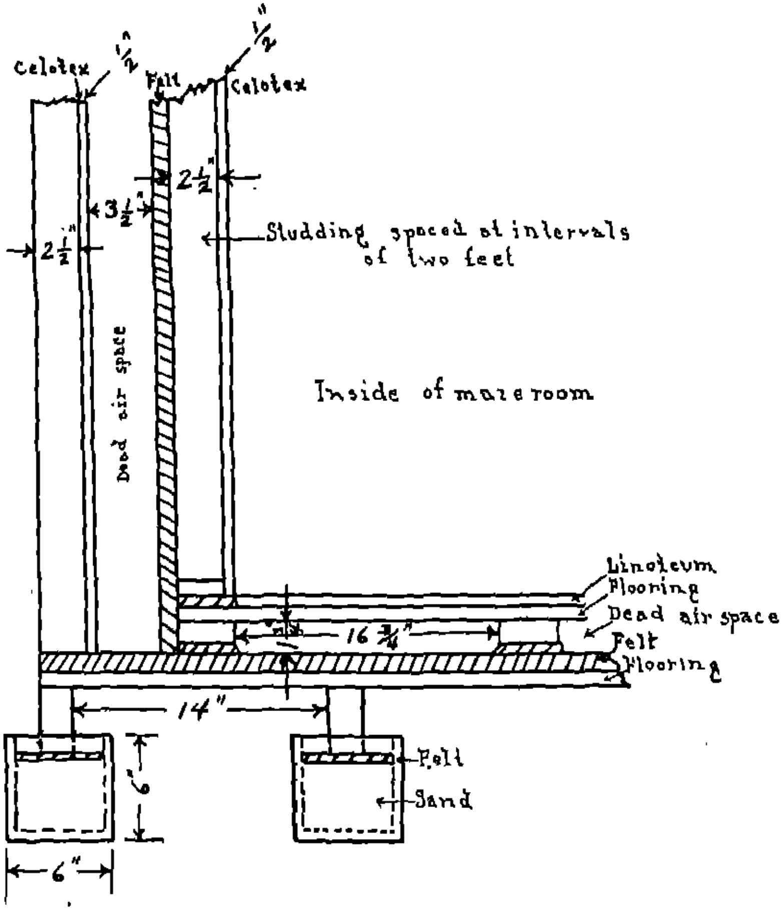

FIGURE 3

Detail of the Construction of the Room for Controlling Sound

½″ in thickness, and of 1″-thick builders' felt. The total structure consisted of two rooms built entirely separate from one another except that the inner one rested on the floor of the outer. The inside dimensions of the inner room were 6′ high by 9′9″ x 9′9″. The floors of these rooms were of matched tongued-and-grooved flooring and were separated by an air space and layers of felt. The entire structure rested in troughs containing felt strips and sand to minimize the vibrations coming from the floor. All the essentials of construction are shown in Figure 3.

Particular care was taken in the construction of the doors, of which there were two, one for each of the two rooms. To make possible the observation of the animals from without the box, each door was constructed with a glass window. The bottom of the inner window was 41″ from the floor of the box. The window in each door was double and was constructed of two panes of glass carefully placed in Celotex or paper felt, and with a dead air space between the two panes. The retracing doors and starting-box door of the maze were operated by wires running through tiny holes in the walls of the rooms. Within the inner room ordinary noises from outside, as from the cages of the rats, could not be heard, and loud sounds could be heard only in a rather muffled way and with their direction confused.

4. *Animals.* The groups of this experiment were relatively homogeneous. Each group was composed of as few litters as were necessary to give the desired number of rats, and in addition to this the different litters were fairly closely related. Most of the animals were

from the Clark Psychology Laboratory colony, which has been developed from Wistar stock and which has been rather generally inbred. The rats of Group C were secured directly from the Wistar Institute.

Table 3 gives the main numerical data on the animals used. A word is in order as to the group symbols used. The group designation is a capital letter (*A, B,* etc.), but the group symbol that will be generally used will indicate also the maze on which trained (I, II, or *El*), the number of retracing doors used (1, 4, or 13), and whether fed liberally or scantily (*Lb.* or *Sc.*). Thus Group I. 13. *Sc.C.* refers to Group *C* as run on Maze I, with 13 retracing doors used, and with a program of scanty feeding. Two of the groups are indicated by *B* and *B'*, respectively, because these two groups were composed of the same litter split as evenly as possible between the two groups in order to make

TABLE 3

Animals Used

Animal groups with their ages, numbers of individuals, numbers by sex, number of trials, and mazes on which used.

Group designation	No. ♂	No. ♀	Total No.	Age in days at start of training on straightaway	No. days between groups of trials	No. trials on maze
I.4.*Lb.A*	23	18	41	85 to 115		30
I.4.*Sc.B*	15	16	31	52 to 102		30
I.1.*Sc.B'*	18	13	31	52 to 102		30
II.1.*Sc.D* I.1.*Sc.D*	12	20	32	110 to 116 180 to 186	40	30 10
I.13.*Sc.C* II.13.*Sc.C*	13	21	34	50 to 51 120 to 121	40	30 20
El.C	11	20	31	150 to 160	0	till learned
I.13.*Sc.E*	13	20	33	111 to 137 166 to 194	43 to 45	7 6

them as comparable as possible in distribution of ability. These two groups were also run at parallel times, half of each litter being started on the 4-door maze and the other half on the 1-door maze at the same time. Thus these two groups were adequately equated not only with respect to heredity and heterogeneity, but also with respect to experimental conditions. In view of the closeness of learning curves and reliability coefficients from these two groups, in contrast with the relationship of other pairs of groups, the following experimental work would be much more conclusive if the split-litter technique had been used throughout.

With Group I.13.*Sc.E,* the interval of 43 to 45 days separates a first and a second group of trials on the *same* maze. The interval consisted of 38 to 40 days of rest, followed by 5 days of training on the straightaway (one trial per day, with the full-length straightaway) before the retest. The procedure was different *with Group C and Group D in that with these groups* the retest was given on a different maze from that used in the original learning. With Group *C* the first group of trials was given on Maze I and the second group of trials on Maze II; with Group *D* the procedure was the reverse of this. With these two groups the interval of 40 days between the learning of the first maze and the learning of the second maze was composed of 30 days of rest and 10 days of training on the straightaway.

For purposes of determining the effect on reliability of varying the amount of retracing permitted, the most important comparisons are of Groups I.4.*Sc.B,* I.1.

Sc.B', II.1.*Sc.D*, I.13. *Sc.C*, and I.13.*Sc.E*. As regards the relation of feeding to reliability, the comparison is primarily between Group I.4.*Lb.A* and Group I.4. *Sc.B*. The test-retest correlations involve Groups I.13. *Sc.C*. and II.13.*Sc.C*, Groups II.1.*Sc.D* and I.1.*Sc.D*, and the first and second groups of trials of Group I.13. *Sc.E*.

5. *Procedure*. During the experiment the rats were fed only after their maze run for the day, the food being a dry powdered McCollum diet. The ingredients (proportion by weights) were:

Whole wheat flour	290
Whole milk powder	75
Casein	40
Salt	6
Calcium carbonate	4
	415

No other food was given throughout the entire period of experimentation except with a few of the very youngest rats which showed inability to keep their weight and strength up sufficiently when shifted to this dry food. Water was before the rats in the living cages all the time; and, when the rats were fed after their daily experimental work, water was available in the food compartments. Prior to the beginning of an experiment, dry food was kept before the animals constantly, and, in addition, their diet was supplemented with bread, milk, and green vegetables.

During the experimental period from 2 to 9 rats were kept in each cage, with the sexes always segregated. Cleaning of cages was done only after the experimenting for the day, and an effort was made to

avoid all other conditions which might disturb the rats during the experimental period.

Care was exercised in controlling the feeding because of the fact that, otherwise, consistent differences in performance might have resulted from differences in hunger. Before the experiment with any group, the rats were weighed once a day for two or three days, to determine the normal weights at that time. A weight schedule was then calculated for each rat on the basis of this normal weight. For example, with some groups the object was to bring the rats to 80% of their normal weight at the end of the first 10 days, and to other percentages at other portions of the learning period. (For the details of this, see Figures 12 and 13 and the accompanying text.) The rats were weighed every day of the experiment and their weight fluctuations compared with those scheduled for them. The weights were controlled by adjusting, for each rat separately, the length of the feeding time in accordance with what seemed necessary to adjust the weights to the schedules. The average feeding time for the scantily fed groups was perhaps about twenty minutes a day. Further data on the weight fluctuation of different groups will be presented later in the section of the results that deals with the relation of weight to learning scores.

With every group, 30 preliminary trials were given on the straightaway before training was begun on the first maze learned (3 trials per day were given for 10 days). Following the tenth day on the straightaway, one trial a day on Maze I or II was given for periods of from 7 to 30 days with different groups. With

Groups *C* and *D* retesting was given after 30 days of rest and 10 days of a second preliminary training. During the 30 days' rest the animals were fed liberally, the diet consisting not only of dry food constantly before them, but also of daily feedings of bread, milk, and green stuff. With the preliminary training for the second maze the same general type of control of feeding was instituted as with the first maze, proper allowance being made for the differences in feeding necessitated by the greater age on the second training. With these retested groups, the second preliminary training consisted of merely one trial per day for ten days on the 6′ straightaway. One trial a day was also given with the second maze learned.

On Mazes I and II all of the rats were run in the afternoon between 1 P.M. and 5 P.M. An effort was made to run the same litters at exactly the same time every day and, in general, there were no greater fluctuations from schedule than about twenty minutes.

The procedure used with the elevated maze was radically different. Only rats of Group *C* were tested with this maze, and the training was given on whatever days the rats of this group finished their training on Maze II. On this elevated maze, training was massed. Trials were given with only 45 seconds intervening between trials until each rat had met the norm of three errorless runs in succession, or four errorless runs in five successive trials. Only a nibble of dry food was allowed at the end of each run. The preliminary training for these trials on the elevated maze consisted of two trials given on a simple elevated straightaway of the

same construction as the regular elevated maze, but with a path 28" long.

Two types of errors have been recorded: (1) *forward errors* when a rat entered a blind alley for a distance of two-thirds or more of its body length when coming to a point of choice from a previous segment of the true path, and (2) *retracing errors* when a rat retraced any segment of the true path, or retraced into a blind. The forward errors rather than retracing errors have been used in the calculations of learning curves, distribution of scores, reliability coefficients, etc., except in the few cases where it is expressly stated that retracing errors have been used. This is the procedure used by Stone and Nyswander (25) and by Heron (6).

The above usage was adopted in the effort to duplicate the procedure of Stone and Nyswander. Recent personal inquiry, however, reveals that the original statement of their definition of types of error was not correctly interpreted. The differences may be illustrated by Figure 4. According to the usage of the present experiment, one forward error, (1), would have been recorded, and six retracing errors, (2), (3), (4), (5), (6), and (7). According to the usage of Stone and Nyswander, two forward errors would have been recorded, (1) and (2), and but two retracing errors—(3), (4), and (5) as the first, and (6) and (7) as the second. Heron's procedure, which also was designed to duplicate Stone and Nyswander's, follows their procedure except that (2) would have been counted as a retracing error (letter from Heron).

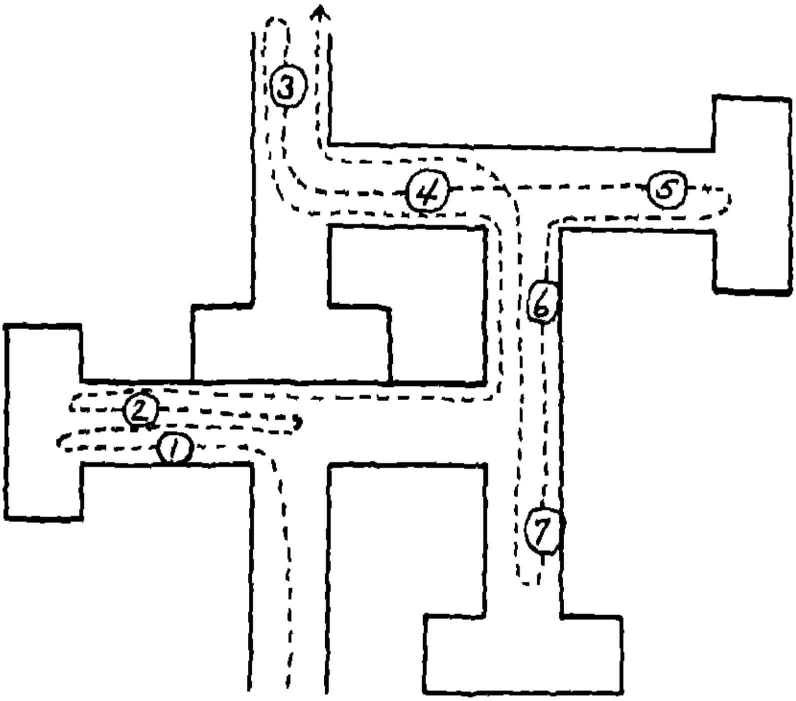

FIGURE 4
DIAGRAM OF THE DIFFERENT TYPES OF ERROR (SEE TEXT)

These differences in definition in these three studies are regrettable, but are not of much practical consequence. The difference in definition of forward error concerns a type of error made only rarely; and, as regards the retracing errors, these have entered into but few of the important calculations.

C. RESULTS

The results of this experiment which are of primary interest are the data on reliability. However, it must

be remembered that reliability coefficients are a function, not merely of the apparatus used, but also of the procedure and of the groups tested. Therefore, rather complete data are presented on the learning curves on the straightaway and maze, on the weight curves, and

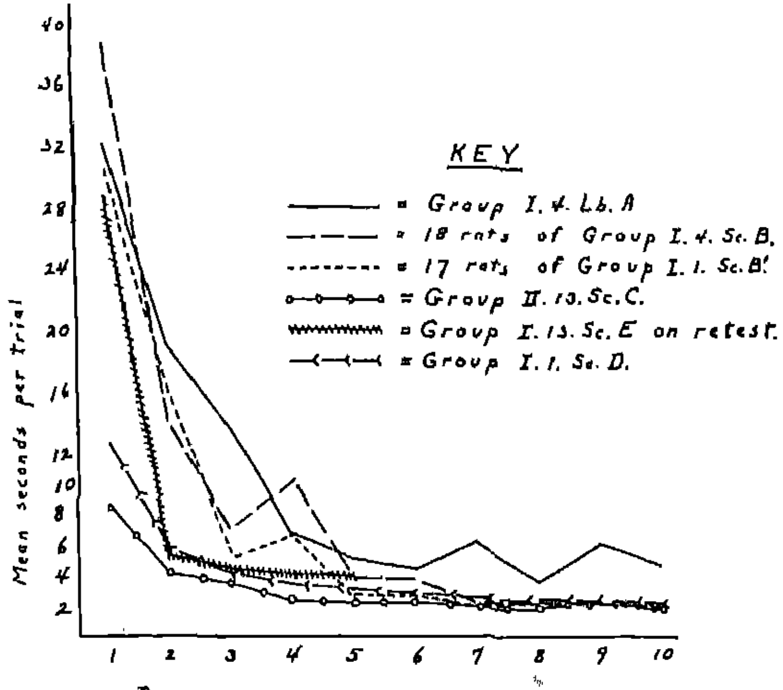

FIGURE 5

MEAN TIME OF THE RUNS ON THE STRAIGHTAWAY OF THOSE GROUPS WITH WHICH THE SIX-FOOT STRAIGHTAWAY WAS USED FOR ALL TRIALS

With Groups II.13.*Sc.C*, I.13.*Sc.E*, and I.1.*Sc.D*, which had been tested before, only one trial a day was given. With Groups I.4.*Lb.A*, I.4.*Sc.B*, and I.1.*Sc.B'*, three trials a day were given, and the value for any particular day represents the mean time of the three runs of that day.

on the variability of scores in different periods of learning.

1. *Performance on the Straightaway During the Preliminary Training.* As indicated on the graphs, certain of the rats had all of their trials on the full-length straightaway, and the remainder were trained for the first three days (9 trials) on a 26" straightaway. It was found that an economy of time was effected by the use of the abbreviated straightaway for these early trials without apparently sacrificing in any way the

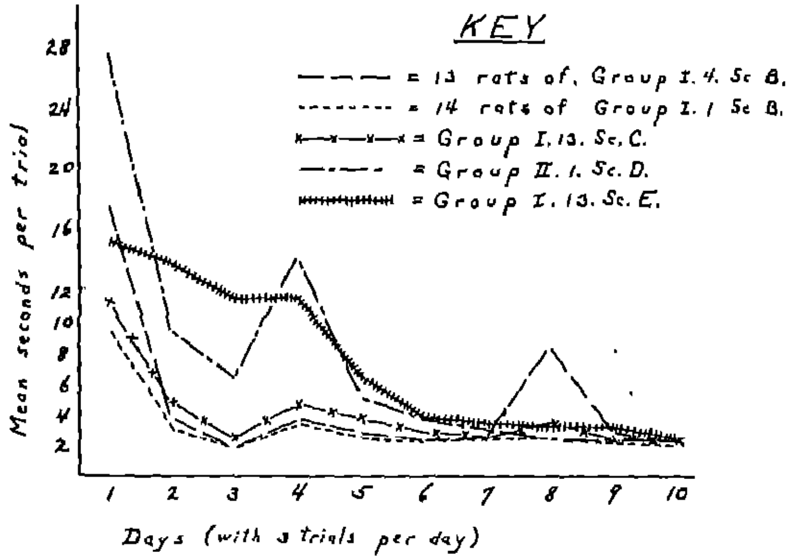

FIGURE 6

MEAN TIME OF THE RUNS ON THE STRAIGHTAWAY OF THOSE GROUPS WITH WHICH THE 26-INCH STRAIGHTAWAY WAS USED FOR THE FIRST THREE DAYS AND THE 6-FOOT STRAIGHTAWAY FOR THE REMAINING SEVEN DAYS OF THE PRELIMINARY TRAINING

The value for any particular day represents the mean time of the three runs of that day.

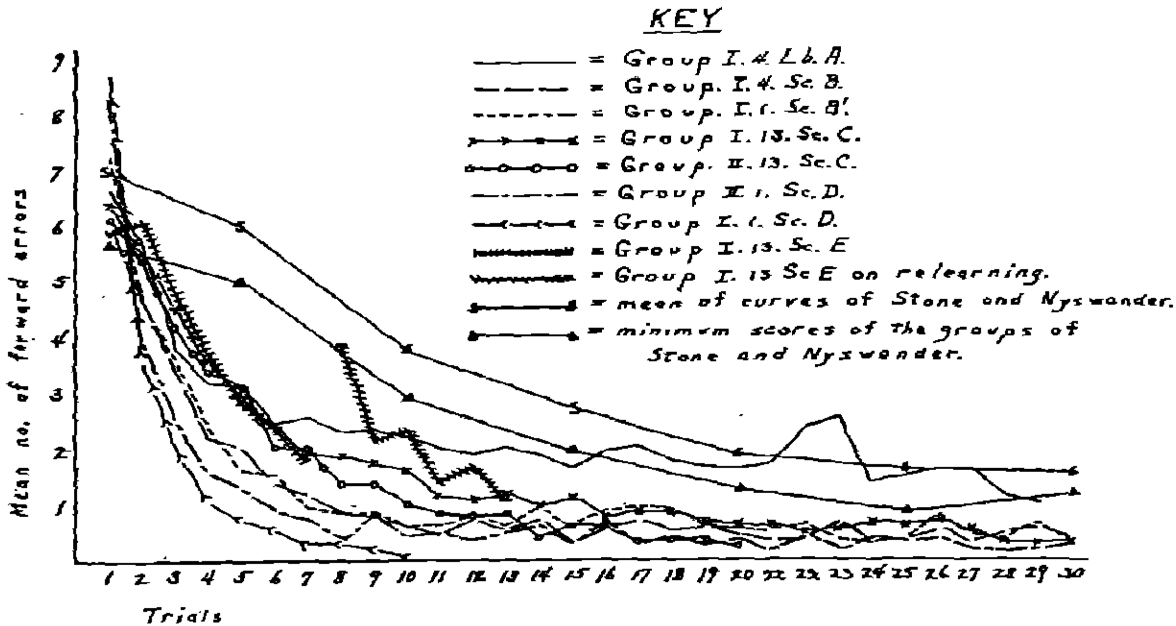

FIGURE 7

Mean Number of Forward Errors per Trial for Different Groups Together with Curves Indicating the Approximate Mean and Minimum Scores of the Groups of Stone and Nyswander (25) on the Same Maze

value of the preliminary training. It, of course, may be that either program of preliminary training may have some slight advantage over the other, but if any conclusions may be drawn from the mean time curves for the later trials on the straightaway (see Figures 5 and 6), it would seem that the difference is very slight indeed, for, from the fifth day on, the mean time curves for the different groups are hardly distinguishable.

2. *Learning Curves of the Different Groups.* The main reason why learning curves of the present study are of interest is the fact that they are markedly lower than any of the learning curves of the groups used by Stone and Nyswander in their study of reliability. This is shown by Figure 7 on which the mean number of errors per trial is graphed for the different groups. To facilitate comparison with the results of Stone and Nyswander, two lines have been added to this figure to represent, first, the mean errors of their eight groups and, second, the minimum values achieved by any of their groups.[4]

It is to be noted that only one of the eight learning curves of the present study has as high values as the lowest of Stone's groups after about the first 3 trials. All the remaining groups of the present study have reached, on an average in 10 trials, a point lower than

[4]Stone and Nyswander in their article on the reliability of the multiple-T maze give no table of the mean errors and, consequently, it has been necessary to draw these graphs from a visual inspection of the graph of the learning curves of their groups. The upper line represents the mean for all eight of their groups as estimated by the apparent mean values for Trials 1, 5, 10, 15, 25, and 30. The lower line indicates the lowest values that any of their groups had on these same trials.

Stone's groups reached on an average in 30 trials. Moreover, whereas the learning curves of Stone and Nyswander fall rather steadily throughout the entire 30 trials, with most groups of the present study, there is practically no improvement after the first 10 trials. It is probable that this difference will be reflected to some extent in the reliability coefficients, because the situation in the present experiment would seem to correspond to the situation where an educational test is applied to a group for which (with the exception of a few very difficult items) the material is too simple. The accumulation of scores at zero tends to give lower reliability coefficients for such a group, other things being equal, than would be secured with the same test with a group having a range of ability such as would tend to give normal distribution of scores.

There are a number of differences between the present experiment and Stone and Nyswander's which may be responsible for the differences in rapidity of learning. These differences are: (*a*) use of the room to control sound, (*b*) preliminary training of 10 days rather than of 5, (*c*) use of the straightaway in preliminary work rather than the problem escape box used by Stone and Nyswander, (*d*) the 4" greater length of units in the present maze, (*e*) possible differences in strength in motivation, and (*f*) differences in the animals.

It would not seem probable that differences in motivation are chiefly responsible for the difference in learning unless Stone and Nyswander are mistaken in their conclusion that they had found and used the feed-

ing schedule which provided approximately optimum conditions for rapid learning.

It is to be noticed that the learning curves of the present study, in spite of their essential similarity to most of the learning curves, tend to fall more or less into several distinct groups. Thus, the 13-door groups (I.13.*Sc.C*, II.13.*Sc.C*, and I.13.*Sc.E*) have somewhat the highest learning curves (except for Group I.4.*Lb.A* in the latter portions of its course). The 1-door groups (II.1.*Sc.D*, I.1.*Sc.D*, and I.1.*Sc.B'*) have, in general, the lowest learning curves. Group I.4.*Sc. B* and Group I.1.*Sc.B'*, which, it will be remembered, were formed of split litters and run on the same days, have virtually the same learning curves. Group I.4.*Lb.A*, after the sixth trial, has a learning curve much higher than any other group. It will be shown later that the explanation for this is the different feeding program used with this group, and, hence, considering merely the groups with roughly equal motivation, it can be seen that the graphs of Figure 7 indicate an inverse relationship between the number of retracing doors used and the rapidity with which forward-going errors are eliminated.

The mean retracing errors for the various groups for the early trials are shown in Figure 8. It is to be noted that the curves drop much more sharply than do the curves for forward errors (see Figure 7). Most of the retracing errors were made on the first trial alone, and after the fifth or sixth trial the only group that made an appreciable number of retracing errors was Group I.4.*Lb.A*. From the fact that retracing errors

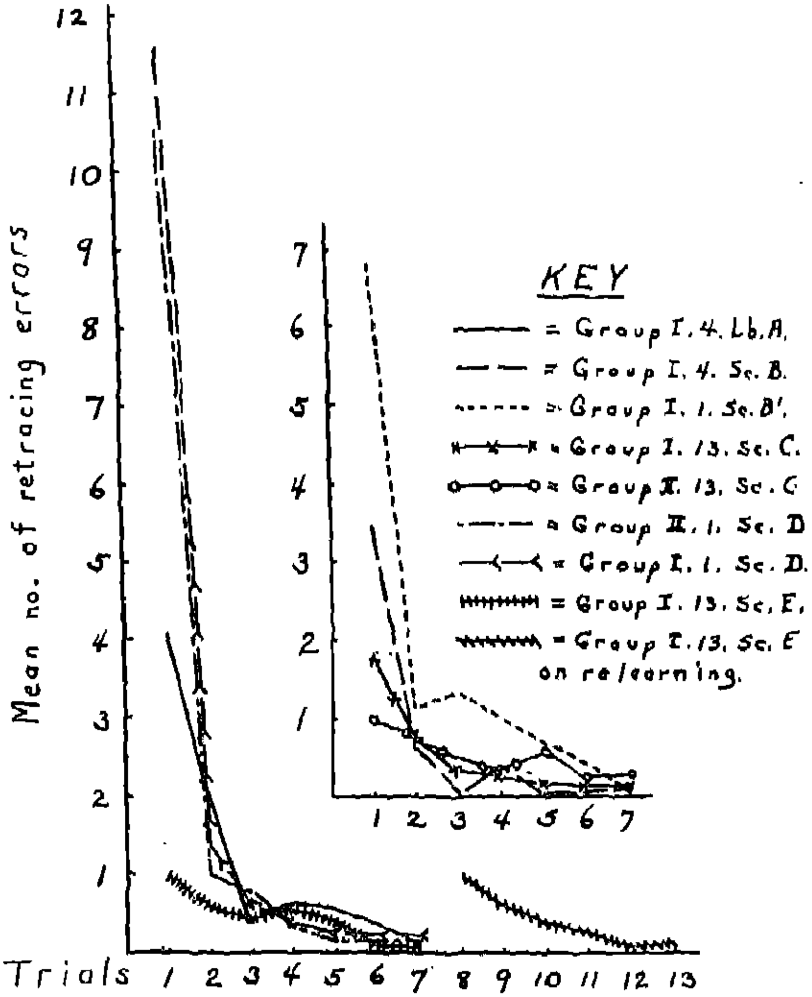

FIGURE 8

Mean Number of Retracing Errors per Trial for the Different Groups in the Early Period of Training

In the trials subsequent to the seventh, all of the groups retain about the same level as reached at the seventh trial, except for Group I.13. *Sc.C*, as shown in the figure (with this group a period of 43 to 45 days separates Trials 7 and 8), and Group I.4.*Lb.A*, the curve of which remained at about .5 retracing error per trial throughout—about double or triple the level of any other group.

are concentrated so largely in this first trial, and from the fact that this first trial is so largely a matter of chance, it justly seems that maze measures will be improved either by disregarding retracing errors altogether or else by disregarding the first trial, as Tolman, particularly, has recommended. That either of these courses, or even the two of them together, would be only a partial remedy, however, is indicated by the fact, as pointed out in connection with Figure 7, that retracing errors on the first trial seem to be significantly related to the forward errors made even after the first trial. That is, there seems to exist a direct relationship between the number of retracing errors made on the first trial and the rapidity with which forward errors are eliminated on the subsequent trials. This furnishes a strong argument in favor of the maximum prevention of retracing.

The same general relationships are brought out by the data on the number of trials required by the different groups to satisfy several norms of learning. The median number of trials of the different groups are shown in Table 4. The severe norm of learning used was the standard of three successive errorless runs, or four errorless runs in five successive runs (with "errorlessness" defined as freedom from either forward or retracing errors). The other, or moderate norm, required three successive runs with a maximum of but one forward error (retracing errors disregarded). The scores are in terms of number of trials *preceding* these errorless runs. Group I.13.*Sc.E* is not included in this table because of the fact that in the

TABLE 4

MEDIAN NUMBER OF TRIALS REQUIRED BY DIFFERENT GROUPS BEFORE MEETING THE NORMS OF LEARNING

Group	Median number of trials preceding the moderate norm	Median number of trials preceding the severe norm
I.13.*Sc.C*	13	17
II.13.*Sc.C*	9	10
I.4.*Lb.A*	23	31
I.4.*Sc.B*	8	10
I.1.*Sc.B'*	7	8
II.1.*Sc.D*	6	8
I.1.*Sc.D*	4	5
El.C.	7	

13 runs given this group only 39% of the group satisfied the moderate norm and only 24% the severe norm. With Group I.4.*Lb*.A only 49% of the group satisfied the severe norm in the 30 trials scheduled for the group, but additional trials were given to determine the median.

The learning curves in terms of time[5] are shown in Figure 9. Attention is called to the following points. First, even with the groups with which retracing was permitted the greatest mean time for the first trial was less than four minutes. This was the case without the discarding of a single animal, and it is a strong recommendation for the type of preliminary training used in the present experiment. Secondly, there is a close similarity between the split-litter groups, I.4.*Sc.B* and I.1.*Sc.B'*. Closer correspondence exists between the

[5]Taken from point of leaving the spot marked *S* on the diagram of the maze in Figure 1 to the instant of passing the door of the food box, with no omission of time for stops in the maze. Very few stops occurred, so that the time corresponds closely to actual running time anyway.

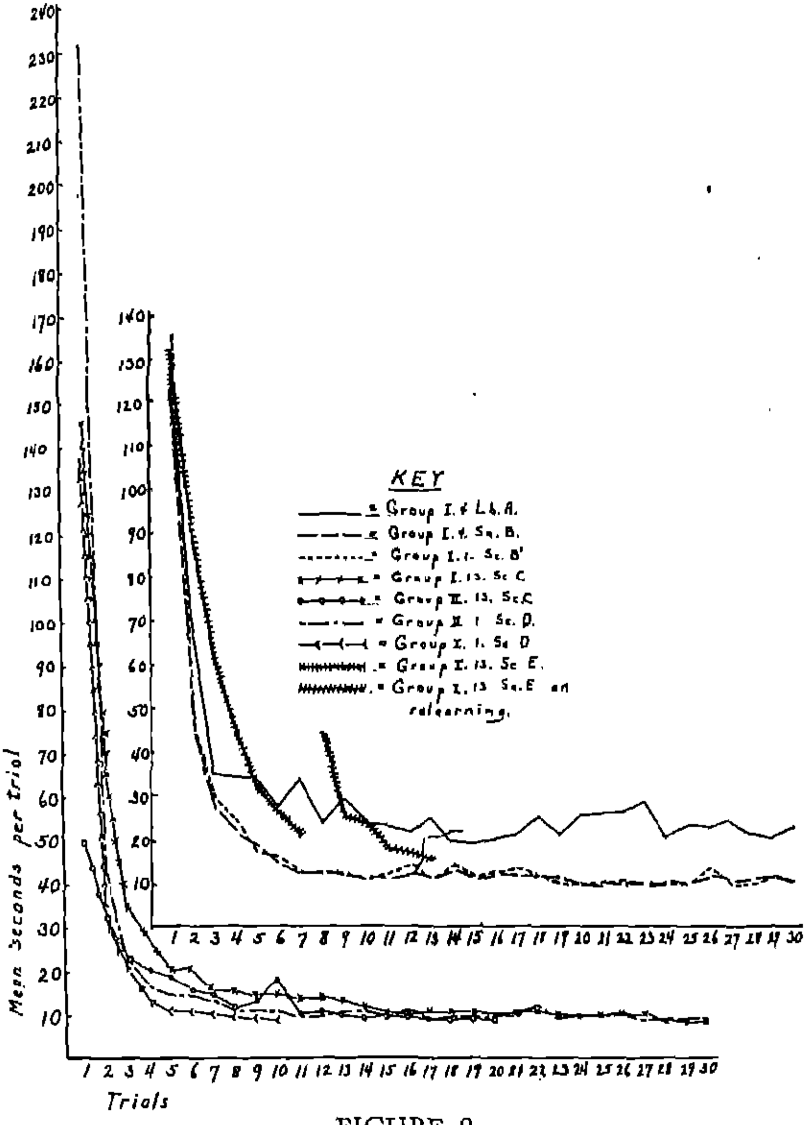

FIGURE 9

MEAN NUMBER OF SECONDS PER TRIAL FOR THE DIFFERENT GROUPS

time curves for these two groups than between the time curves for any two other groups.

As regards transfer from the first to the second maze learned, there is relatively little evidence of this in the learning curves for errors, but there is very definite evidence, in the time scores of the first trial on retest, of transfer with respect to speed of running.

3. *Distribution of Scores in Different Periods of Learning.* The data on the distribution of scores in different periods of learning are an important supplement to the learning curve. Liggett (18), for instance, has pointed out the relation of skewness to the problem of the choice of the best measure of central tendency,

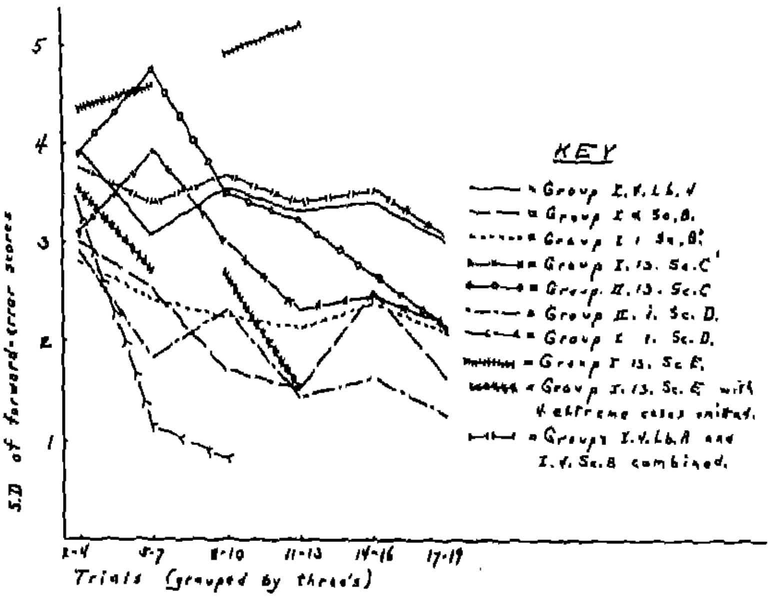

FIGURE 10

STANDARD DEVIATIONS OF FORWARD-ERROR SCORES FOR DIFFERENT GROUPS OF THREE TRIALS

and in the theoretical section of the present paper, in the discussion of the relation between range of ability and the reliability coefficients, a number of problems were discussed which can be illuminated to some extent now by the data on distribution of scores in the present study.

The sigmas of scores in different periods of learning may first be considered. Figure 10 and Table 5 show the sigmas of forward-error scores in the different groups of the three trials. It can be seen from a comparison with Figure 7 that there is a tendency for the degree of scatter to vary in proportion to the height of the learning curve. Thus Group II.1.*Sc.D*, which has the lowest learning curve, also tends to have the smallest sigmas. Groups I.4.*Sc.B* and I.1.*Sc.B′* have the next largest sigmas, as well as the next highest learning

TABLE 5

VARIABILITY OF FORWARD-ERROR SCORES AT DIFFERENT PERIODS OF THE LEARNING PERIOD AS INDICATED BY THE SIGMAS OF FORWARD-ERROR SCORES FOR THE DIFFERENT GROUPS IN VARIOUS GROUPS OF TRIALS

	Trials						
Group	1	2-4	5-7	8-10	11-13	14-16	17-19
I.13.*Sc.C*	1.75	3.09	3.94	3.00	2.34	2.48	1.91
II.13.*Sc.C*		3.88	4.75	3.48	3.22	2.68	1.89
I.13.*Sc.E*	1.59	4.34	4.57	4.93	5.19		
I.13.*Sc.E* with 4 extreme rats dropped		3.55	2.72	2.71	1.49		
I.4.*Lb.A*		3.97	3.10	3.54	3.33	3.40	3.02
I.4.*Sc.B*	1.98	3.03	2.55	1.72	1.52	2.49	1.62
I.4.*Lb.A* and I.4.*Sc.B* combined		3.78	3.38	3.67	3.40	3.53	3.05
I.1.*Sc.B′*	2.63	2.81	2.46	2.26	2.13	2.38	2.10
II.1.*Sc.D*	3.84	2.94	1.83	2.07	1.43	1.62	1.25
I.1.*Sc.D*	6.34	3.49	1.17	.83			

curve, and Groups I.13.*Sc.C*, II.13.*Sc.C*, and I.13.*Sc.E*, and I.4.*Lb.A* have the largest sigmas and highest learning curves.

One line on Figure 10 demands comment, namely, the line for "Groups I.4.*Lb.A* and I.4.*Sc.B* combined." These two groups have been combined in certain calculations in order to illustrate how reliability coefficients are affected by having unequal feeding for different members of a group. These groups were run under similar conditions except for the fact that Group I.4.*Lb.A* was fed much more liberally than the other group. The sigmas of the combined group are seen to be only slightly higher than the sigmas for the more variable group (Group I.4.*Lb.A*), but, as will be brought out later, the reliability coefficients in every case but one are quite appreciably higher than for either group treated separately. *These data, accordingly, are a good illustration of the point that only the most careful control of experimental conditions can yield reliability coefficients which are not affected by systematic errors.*

The above table and figure also include the data for Group I.13.*Sc.E* with four extreme cases dropped from it, in order to show what effect a few extreme cases can have. It will be shown later that dropping these same four rats from the correlations of this group affect the correlation coefficients as seriously as they affect the sigmas in this case (although still not altering the main conclusions drawn from the experiment). Such cases as these offer a rather baffling problem to the experimenter, whether he is working on the problem of re-

liability or on the question of the influence on learning of some experimental factor. My own judgment would be that where a few individuals deviate markedly from the group (say, by three sigmas), that the experimenter should conclude that special factors must be operating in their case, whether he could specify what those special factors were or not, and hence should exclude them from his data. (In this instance I had been hesitant, beforehand, about using these rats, as the litter had not seemed to be in the best physical condition. However, the other two rats of the litter made almost the best records of the group.)

As regards the skewness of maze data, and the tendency for this skewness to increase the longer the training is continued, the results graphed in Figure 11 may be taken as typical. Group I.4.*Lb.A* is the group with least tendency to skewness of all groups, and Group I.4.*Sc.B* is typical of all the remaining groups except for the fact that the distributions of scores in the remaining groups (except I.13.*Sc.E*) are even slightly more markedly skewed in the later periods of learning than are the distributions of Group I.4.*Sc.B*.

The significance of these data is not to be underestimated relative to the problem of reliability. Altogether apart from the fact that such accumulations of zero scores very possibly affect reliability coefficients, a maze procedure which yields such an accumulation of zero scores after so few trials might quite reasonably be condemned merely for this one reason. Other things being equal, group differences will be more clearly brought out by measuring instruments which do not

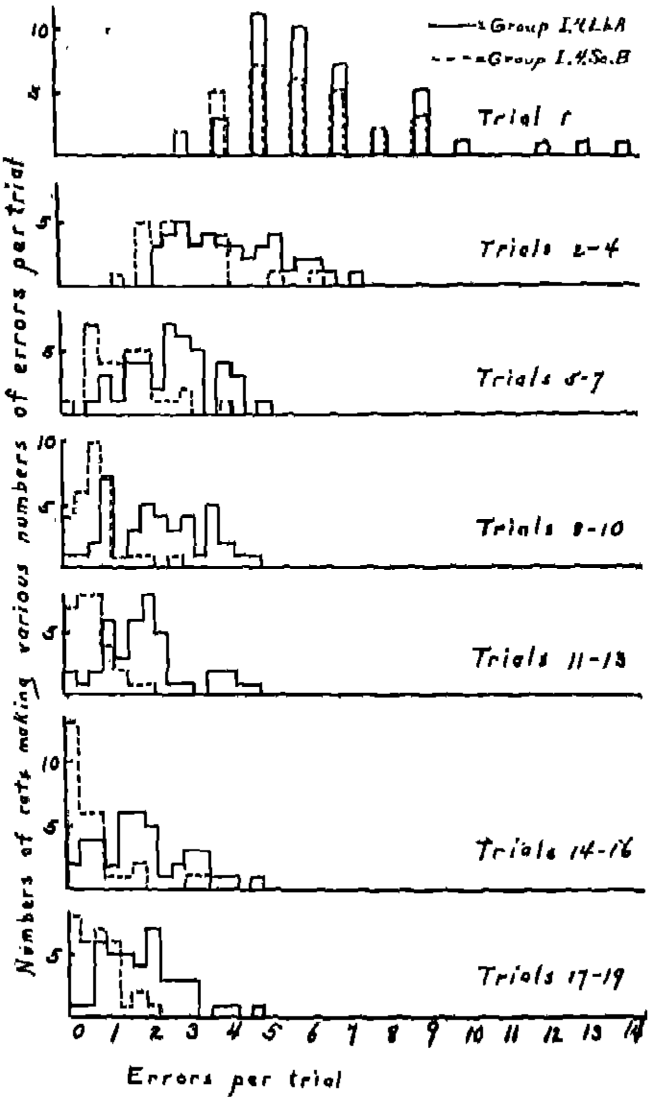

FIGURE 11

Distribution of Forward-Error Scores in Successive Periods of Training with Groups I.4.*Lb.A* ($N=41$) and I.4.*Sc.B* ($N=31$)

have an accumulation of zero scores, because on such a measuring instrument the differences of ability of the entire group will find expression, whereas with scores accumulating at zero, one no longer has any differential measure of a good part of the group. Therefore, it would seem quite probable that the reliability of the multiple-T maze would be increased if it were made considerably longer and more difficult. It would, of course, be possible with the 12-unit multiple-T maze to secure data which would not be markedly skewed if one dealt merely with the first seven or eight trials (with the experimental conditions the same as those used in the present experiment). However, to so curtail the length of the training period would shorten the test in a way that might be expected to lower the reliability. With a longer and more difficult maze one might have the greater length of test which is so favorable to reliability, and at the same time not be compelled to utilize markedly skewed data.

4. *Correlations between Scores on the First Trial and Scores on Subsequent Trials.* Another problem closely related to that discussed above is the problem of the correlation between performance on the first trial (which, as we have seen, is markedly influenced by the number of retracing doors used) and scores on the subsequent trials. These correlations are of particular interest because of the light they throw on the suggestion that has been made, particularly by Tolman and Tryon, that the data for the first trial or two should be discarded.

TABLE 6

CORRELATIONS BETWEEN SCORES ON THE FIRST TRIAL AND SCORES ON SUBSEQUENT TRIALS

Group	Trials involved in the correlations	Correlations of forward errors	Correlations of time scores
I.13.*Sc.C*	1 *vs.* 2-30		.52±.13*
II.13.*Sc.C*	1 *vs.* 2-20		.06±.17
I.13.*Sc.E*	1 *vs.* 2-7	.22±.17	
I.4.*Lb.A*	1 *vs.* 2-30		.23±.15
I.4.*Sc.B*	1 *vs.* 2-30	.14±.18	.18±.17
I.1.*Sc.B'*	1 *vs.* 2-30	—.02±.18	—.06±.18
II.1.*Sc.D*	1 *vs.* 2-30	—.32±.16†	.00±.18
I.1.*Sc.D*	1 *vs.* 2-10	—.42±.15	—.21±.17

*In this case the size of the correlation, with one extreme case eliminated, is .23±.17, which is more descriptive of the tendency with the group as a whole.

†When these same trials are used, but with *all* errors on Trial 1 *vs.* forward errors on Trials 2-30, the correlation is —.35±.16.

The correlations which have been made[7] (see Table 6) include all of the groups where there was reason to suspect that the first run might have had some differential effect among the members of a group. In this series of correlations the only figures which would seem to indicate any probable relationships are the correlations for Group *D* (II.1.*Sc.D* and I.1.*Sc.D*). It will be remembered that this group had the lowest learning curve of all the groups, except on the first trial. It may be, therefore, that with this group conditions were more nearly optimal for learning than with Group I.1.*Sc.B'*, which was also trained with but one retracing door, and that for this reason appreciable negative correlations were found with it, but not with Group I.1.*Sc.B'*. These

[7]In all of the original calculations of this paper the standard error is used in preference to the probable error, so that henceforth in this paper $r=.22\pm.17$, for instance, is to be read as "r equals .22, with a standard error of .17."

negative correlations indicate that very probably with some groups the learning which goes on in the first trial is not a constant for all rats and therefore cannot be discarded without eliminating some of the significant data. When this is considered relative to the fact that scores on this first trial are largely determined by chance factors, and that their sigmas are so very large that inclusion of the scores for the first day may be expected certainly to obscure the true values in an unusual degree, we have a strong argument in favor of that type of maze with which the first trial will be approximately a constant as far as its effect on learning is concerned. It is very probable that this type of maze is secured by the use of doors to prevent retracing.

5. *Relation of Weight Changes to Learning Curves and Variability.* Various studies have shown that speed of maze learning is related to strength of motivation. Hence, to understand the reasons for the different learning curves secured in the present experiment, it is important to have a careful analysis of the weight changes. The data on weight changes are presented in two different ways in Figures 12 and 13. In Figure 12 the weight curves are plotted on a *semi-logarithmic chart,*[8] along with the curves of normal growth in weight de-

[8]The advantages of this system of plotting on a *semi-logarithmic chart* are that equal vertical distances at any point on the chart represent equal percentages of gain or loss, and lines of corresponding slope anywhere on the graph indicate that weight is falling off or increasing at the same percent in each case. This is a distinct advantage for the inspection of weight curves because of the fact that it is the same percentage change in weight which will make two groups equal in motivation rather than the same absolute change in weight, in case the original weights are different.

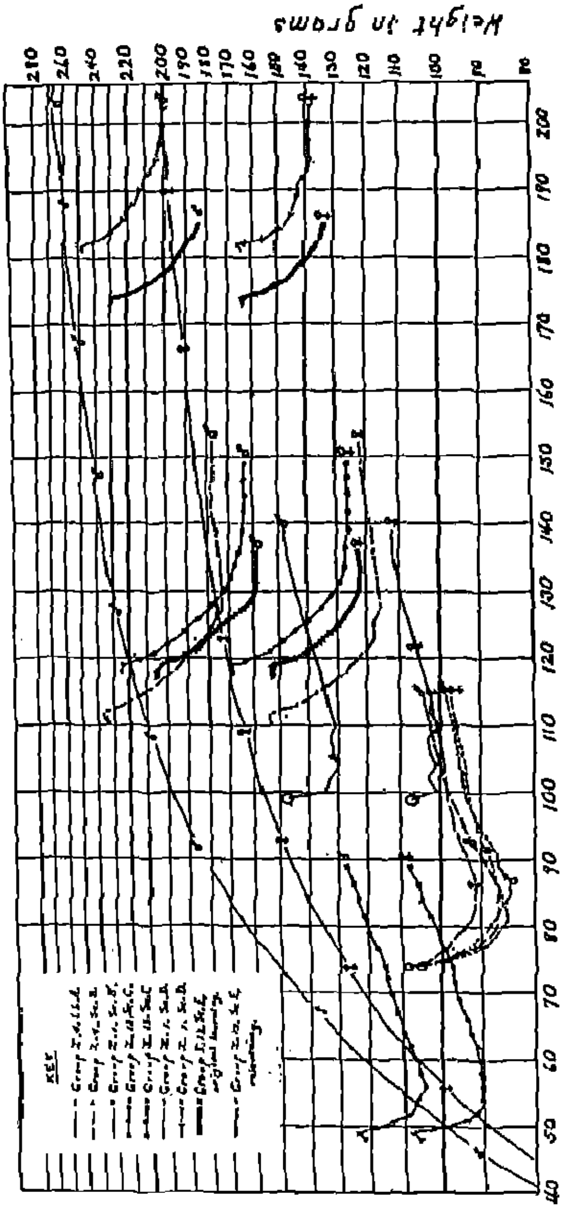

FIGURE 12

MEAN WEIGHTS OF THE MALE AND FEMALE RATS OF THE DIFFERENT GROUPS DURING THE PERIODS OF STRAIGHTAWAY AND MAZE TRAINING, TOGETHER WITH THE CURVE OF NORMAL WEIGHT AT DIFFERENT AGES

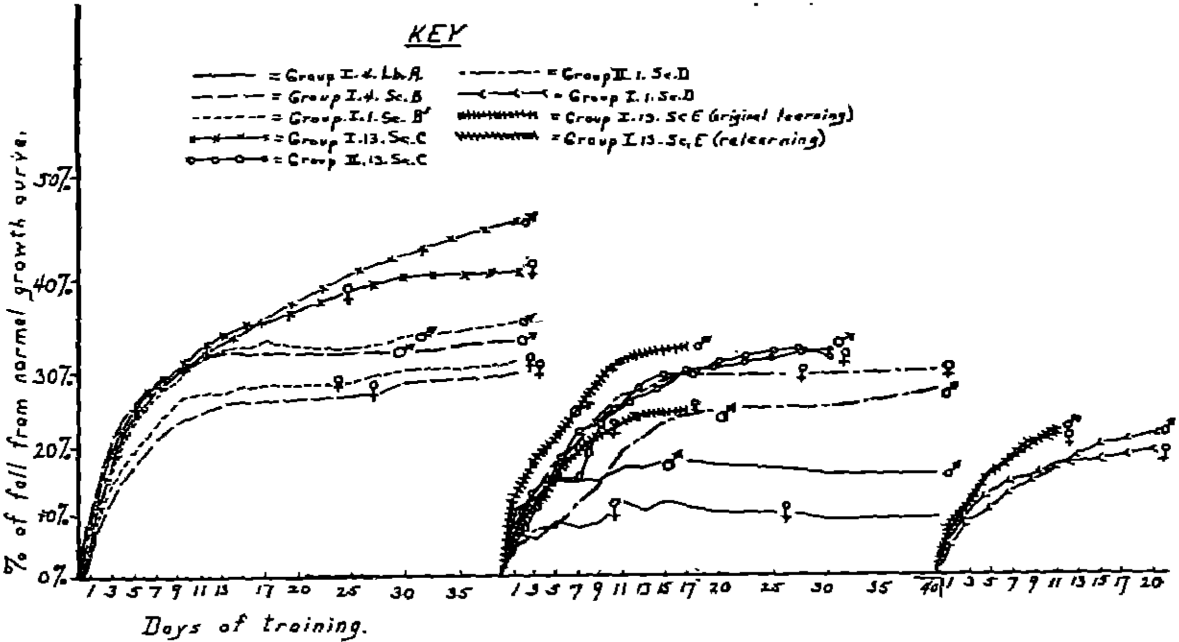

FIGURE 13

Percentage of Deviation of Mean Weight Curves for the Various Groups from the Curves of Normal Weight Change for Their Ages

The groups of approximately the same ages are grouped together.

rived from King (13). It is essential to consider weight changes not merely in terms of percentage of gain or loss over a certain period, but also by comparison with the rate of growth that would probably have occurred with the same group under normal conditions. For example, the fact that the weight curve for Group I.13.*Sc.C* was rising after the first 10 days does not mean that this group was necessarily less strongly motivated on this maze than on Maze II where the weight curve fell steadily until almost the end of the training, inasmuch as the two training periods fell in the periods of rapid and slow growth, respectively.

In order to facilitate the comparison of weight changes in terms of their deviation from normal growth, Figure 13 has been added to supplement Figure 12. Figure 13 shows the relative deviation of the different groups from the curve of growth which they probably would have had under normal feeding, judging on the basis of the growth curves furnished by King. In other words, Figure 13 shows the relative degree to which the weights of different groups were reduced in comparison with normal weights, and it is probably safe to say that the heights of the different curves of this graph represent roughly the relative strength of motivation with the different groups. Comparison of this graph with Figures 7, 8, and 9 reveals that Group I.4.*Lb.A* is the only one in which the rate of learning was mainly determined by the feeding program used. This was the group with which weight was least reduced. As regards the other groups, the amount of retracing permitted seems to have been more influential in de-

termining the rate of learning than was the feeding program used. Thus, of the youngest groups, I.13.*Sc.C* had its weight cut more than did either Group I.4.*Sc.B* or Group I.1.*Sc.B'*, and yet the learning of Group I.13. *Sc.C* did not proceed as rapidly. Similarly, with the two oldest groups, I.13.*Sc.E* and I.1.*Sc.D,* the learning was more rapid with the group permitted to retrace than with the group with the greater weight drop but with retracing prevented. Of the remaining groups, the most important comparison is between Group I.13.*Sc.E.* and II.1.*Sc.D,* and here likewise the same situation is found.

One important aspect of the control of weight, however, cannot be deduced from an inspection of curves of means. We want to know not only what percentage of normal weight the *group* attained at any given point, but also whether the amount of variability which has escaped control affected the speed of learning of different individual rats. Hence, with those groups where there was most suspicion that such might be the case, the correlations were calculated between error scores and loss of weight. The error scores involved in these correlations were the total forward errors on Trials 2 to 10, inclusive, for Groups II.1.*Sc.D* and II.13.*Sc.C,* and for Group I.13.*Sc.E* those on Trials 2 to 7 in one case and on trials 8 to 13 in another. (These trials were selected, in the case of the first two groups mentioned, because of the fact that they covered the period of most rapid learning and also probably represent the most reliable period of learning.) The weight figure correlated was the percentage of the origi-

nal weight to which the weight had fallen by the fourth trial on the maze with each individual animal. The figure for weight at Trial 4 was estimated by averaging the weights on Trials 3, 4, and 5 in order to eliminate chance effects of daily fluctuations in weight. The resulting correlations are given in Table 7.

A negative correlation in these cases would mean that the rats whose weights had been cut more drastically tended to make larger scores, and vice versa. The only one of the correlations which appears possibly significant is the correlation for Group II.1.*Sc.D*. However, the correlation in this case is largely the result of the influence of a single extreme case, whose weight was not extreme in that it had been reduced to only 70% of the original weight rather than to 75%, as was the case with the group mean, but whose error score was over six sigmas away from the mean of the group. With this one rat omitted from the correlation, the figure for Group II.1.*Sc.D* is $r = -.14 \pm .18$. This figure is prob-

TABLE 7

CORRELATIONS BETWEEN PERCENTAGE OF ORIGINAL WEIGHT (W) AND ERRORS ON CERTAIN TRIALS (TRIALS INVOLVED INDICATED BY SUBSCRIPTS TO E)

Group II.13.*Sc.C*	$r_{W.E_{2\text{-}10}} = +.28 \pm .16$
Group II.1.*Sc.D*	$r_{W.E_{2\text{-}10}} = -.39 \pm .15$
Group I.13.*Sc.E*	$r_{W.E_{2\text{-}7}} = -.19 \pm .17$
Group I.13.*Sc.E*	$r_{W.E_{8\text{-}18}} = -.09 \pm .17$

(With Group I.13.*Sc.E*, the correlation between percentage loss of weight on learning and percentage loss of weight on relearning was $.20 \pm .17$.)

ably more indicative of the true relationship (or lack of relationship) which tended to exist between feeding and errors within different groups than is the correlation with this rat included.

It therefore seems safe to conclude that the weight control used in this experiment was exact enough to guard against such differences in feeding (within a given group) as might have resulted in consistent differences in maze performance.

6. *Reliability Coefficients from the Present Experiment.* Having presented the auxiliary material which is necessary to facilitate and safeguard the interpretation of the reliability coefficients of the present investigation, we may now turn to a consideration of those coefficients themselves. Four methods of calculating reliability coefficients have been used: (1) correlation of scores on odd and even trials for different groups of trials; (2) correlation of scores on the first, second, and third groups of ten trials, etc.; (3) correlation of errors in various groups of three trials from the second to the nineteenth trial inclusive; and (4) correlation of scores on test and retest (with the same maze used on test and retest with one group, and with a mirror image of the first maze used in the retest with two other groups).

Where the reliability coefficients have been derived from two groups of trials which really constituted a unit of training (such as coefficients from odd *vs.* even trials or from Trials 1-10 or 11-20, for instance) the correlation coefficients could have been corrected by the Brown-Spearman formula for halving of the data:

$$R = \frac{2r}{1+r}$$

(with the standard errors of these calculated by the formula provided by Shen (22)

$$\sigma_R = \frac{2(1-R)}{\sqrt{N}})$$

following the procedure of Stone and Nyswander. However, as indicated before, I believe that the correction would not be warranted because of the fact that *the two halves of maze data are not sufficiently independent to satisfy the necessary conditions for the use of this formula.*

Moreover, there is always a constant relation between r and R, so that the same relations existing among the R's would exist also in the uncorrected r's, and anyone who wished the R's could readily calculate them. In general, the relationship runs thus:

r	R	r	R
.10	.18	.55	.71
.15	.26	.60	.75
.20	.33	.65	.79
.25	.40	.70	.82
.30	.46	.75	.86
.35	.52	.80	.89
.40	.57	.85	.92
.45	.62	.90	.95
.50	.67	.95	.97

In evaluating the reliability coefficients from the present study, most weight should be attached to the coefficients from approximately the first 10 trials because, as may be seen from Figures 7 and 9, with most of the groups the learning curve did not drop appreciably after the first 10 trials. The correlations of trials

after the first 10, therefore, may be fittingly spoken of as correlations of samples of the final level of attainment rather than as correlations of different samples of learning performance. It is interesting to note, however, as will be shown, that even these correlations of errors on the later trials yield fairly high coefficients, both when correlated among themselves and when correlated with the scores on the early portions of learning.

Another reason for attaching greater significance to the reliability coefficients of the early trials is that, if a maze procedure can be found which will give high reliabilities in this portion of the training, that maze procedure will, in general, be more useful for studies of animal learning than another with high reliabilities in some later period, since there is an important economy of time in being able to use the first few trials as a basis for one's experimental conclusions, rather than in having to run the animal for 20 or 30 trials to secure the scores of the latter portions of learning.

a. Correlations of odd vs. even trials. The reliability coefficients from correlations of odd versus even trials may be presented first. Their values are indicated in Tables 8 and 9.

In the case of forward-error scores in the first 10 trials the three groups run with 13 doors had the highest reliability coefficients; the two groups with 4 doors, the next to lowest coefficients; and the two groups with one door, the lowest coefficients. The same relationship exists for odd *vs.* even trials in the first 20 trials and in the first 30 trials, except that one of the 1-door groups sur-

passes one of the 4-door groups. For Trials 11-20 relatively little difference exists between the different reliability coefficients, except for Group II.1.*Sc.D*. For Trials 21-30 the picture changes markedly from any of the previous ones, except that the 13-door group is still highest, and one of the 4-door groups and one of the 1-door groups hold the intermediate values.

In the case of reliability coefficients from the time scores, the same general relationship holds, except that

TABLE 8

RELIABILITY COEFFICIENTS FROM ERROR SCORES ON ODD *vs.* EVEN TRIALS IN DIFFERENT PORTIONS OF TRAINING

(The figures in parentheses are rank-difference correlations; the other figures are product-moment correlations and their sigmas.)

Group	Correlations from odd *vs.* even trials in different portions of training 1-10	11-20	21-30	1-20	1-30
I.13.*Sc.C*	.72±.08 (.70)	.70±.09 (.66)	.86±.04	.83±.05 (.82)	.88±.04
II.13.*Sc.C*	.74±.08	.90±.03		.88±.04	
I.13.*Sc.E**	.78±.07 (.61)	.88±.04			
I.4.*Lb.A*	.37±.13 (.34)	.62±.10 (.63)	.61±.10	.73±.07 (.71)	.74±.07
I.4.*Sc.B*	.60±.11 (.53)	.66±.10 (.51)	.17±.17	.58±.12 (.49)	.61±.11
I.1.*Sc.B'*	.27±.17 (.28)	.71±.09 (.66)	.53±.13	.60±.11 (.51)	.65±.10
II.1.*Sc.D*	.14±.17 (.31)	.28±.16 (.16)	.10±.18	.37±.15 (.07)	.46±.14
I.1.*Sc.D*	—.44±.14				

*With Group I.13.*Sc.E* the first correlation is of the odd and even trials in *Trials 2-7*, the second correlation is for Trials 8-13. With the four extreme cases dropped these correlations become .57±.11, and .34±.16, respectively.

TABLE 9

RELIABILTY COEFFICIENTS FROM TIME SCORES ON ODD *vs.* EVEN TRIALS IN DIFFERENT PORTIONS OF TRAINING

	Correlations from odd *vs.* even trials in different portions of training					
Group	1-10	11-20	21-30	1-20	1-30	All trials except the first
I.13.*Sc.C*	.77±.07	.68±.09	.72±.08	.80±.06	80.±.06	.79±.07
II.13.*Sc.C*	.50±.13	.85±.05		.76±.07		.84±.05
I.13.*Sc.E**	.81±.06	.77±.07				
I.4.*Lb.A*	.57±.11	.79±.06	.80±.06	.74±.07	.86±.04	.86±.04
I.4.*Sc.B*	.33±.16	.77±.07	.50±.13	.41±.15	.59±.12	.81±.06
I.1.*Sc.B'*	—.04±.18	.41±.15	.46±.15	.10±.18	.20±.17	.51±.13
II.1.*Sc.D*	—.20±.17	.10±.18	.15±.17	.02±.18	.03±.18	.29±.16
I.1.*Sc.D*	—.14±.17					.21±.17

*The first of the two correlations for Group I.13.*Sc.E* is for Trials 2-7, and the second for Trials 8-13. With the four extreme cases of this group dropped, the two correlations become .29±.16, and .76±.07, respectively.

the 1-door groups have definitely lower reliability coefficients than with the error scores, while the time-score reliability coefficients of the other groups are about the same. Also, it may be noted that with the three groups with most rapidly falling learning curves (Groups II.1.*Sc.D*, I.4.*Sc.B*, and I.1.*Sc.B'*) dropping the first trial from the calculations decidedly improves the reliability of time scores.

b. Correlations of groups of ten trials and of fifteen trials. The reliability coefficients from correlations of scores on the different groups of 10 or 15 trials may next be considered. Their values are indicated in Tables 10 and 11.

In the case of the error scores, more similarity between the groups exists with the reliability coefficients derived in this manner than between those derived in any other manner. Particularly in the correlations be-

TABLE 10

RELIABILITY COEFFICIENTS FROM FORWARD ERRORS IN DIFFERENT GROUPS OF TEN AND FIFTEEN TRIALS

	Correlations from different groups of trials			
Group	1-10 *vs.* 11-20	11-20 *vs.* 21-30	1-10 *vs.* 21-30	1-15 *vs.* 16-30
I.13.*Sc.C*	.56±.12	.58±.11	.30±.16	.41±.14
II.13.*Sc.C*	.50±.13			
I.4.*Lb.A*	.37±.13	.56±.10	.18±.15	.43±.13
I.4.*Sc.B*	.34±.16	.59±.12	.51±.13	.65±.10
I.1.*Sc.B′*	.46±.14	.49±.14	.44±.15	.58±.12
II.1.*Sc.D*	.40±.15	.60±.11	—.16±.17	.40±.15

TABLE 11

RELIABILITY COEFFICIENTS FROM TIME SCORES IN DIFFERENT GROUPS OF TRIALS

	Correlations from different groups of trials				
Group	1-10 *vs.* 11-20	11.20 *vs.* 21-30	1-10 *vs.* 21-30	2-10 *vs.* 11-20	2-10 *vs.* 21-30
I.13.*Sc.C*	.22±.16	.46±.14	.17±.17	.27±.16	.38±.15
II.13.*Sc.C*	.40±.14			.39±.15	
I.4.*Lb.A*	.69±.08	.65±.09	.64±.09	.68±.08	.57±.11
I.4.*Sc.B*	.24±.17	.64±.11	.46±.14	.46±.14	.55±.13
I.1.*Sc.B′*	.40±.15	.63±.11	.41±.15	.54±.13	.58±.12
II.1.*Sc.D*	.34±.16	.38±.15	.03±.18	—.21±.17	—.20±.17

tween the second and third group of 10 trials, and to some degree between the first and second group of 10 trials, the coefficients from the different groups are approximately the same. The lowest correlations by this method are for Trials 1-10 *vs.* 21-30 for Group II.1.*Sc.D* and I.4.*Lb.A*. It is not surprising that with Group II.1.*Sc.D* the correlation between Trials 1-10 and 21-30 is so low, because in this group more than in any other there was an extremely marked accumulation of zero scores in the latter portions of the training.

With Group I.4.*Lb.A*, however, the low correlations of 1-10 *vs.* 11-20, 1-10 *vs.* 21-30, and 1-15 *vs.* 16-30 are strong evidence against the reliability of the procedure used with this group. Since Group I.4.*Lb.A* was the one with the high learning curve (secured by more generous feeding than was given any other group), insofar as the distribution of scores was concerned, one would have expected relatively high reliability coefficients with this group. The other group run with four retracing doors (Group I.4.*Sc.B*) had a more restricted feeding, and with this group the highest reliability coefficients in each case are secured, except for Trials 1-10 *vs.* 11-20. It would seem, therefore, that strong, rather than weak, motivation increases the reliability of error scores.

In the correlations of time scores, there seem to be no very consistent and significant differences between the different conditions except for the unexpected excellence of Group I.4.*Lb.A*.

Attention has been called before to the fact that, with most of the groups, the learning curve had virtually met its lowest value by the tenth trial and that scores after that point could more appropriately be spoken of as measuring the final level of attainment achieved than as measuring learning proper. The above correlations, however, indicate that, to an appreciable extent, performance during this final level of attainment is correlated with performance in the period of rapidly decreasing errors. Just why this is the case is not clear from the data; but the problem is important for further investigation.

c. Correlations from groups of three trials. The correlations may next be considered which were calculated from the different groups of three trials. Intercorrelations were calculated between all groups of three trials from 2 to 19 inclusive for all the groups. These are presented in Table 12. It is to be expected that the reliability coefficients calculated in this manner will be somewhat lower than the coefficients calculated from groups of 10 or 15 trials because of the fact that shorter tests, in general, do tend to give lower reliabilities. This system of measuring the internal consistency of

TABLE 12

INTERCORRELATIONS BETWEEN THE FORWARD ERRORS OF DIFFERENT GROUPS OF THREE TRIALS

Group I.13.Sc.C; N=34

Trials	2-4	5-7	8-10	11-13	14-16
5-7	.35				
8-10	.25	.72			
11-13	.27	.53	.59		
14-16	.33	.63	.36	.68	
17-19	.07	.17	.02	.50	.46

Group II.13.Sc.C

Trials	2-4	5-7	8-10	11-13	14-16
5-7	.60				
6-10	.39	.69			
11-13	.32	.46	.77		
14-16	.22	.40	.72	.90	
17-19	.14	.27	.56	.77	.65

Group I.13.Sc.E (with 38 to 40 days of rest and 5 days' straightaway training between trials 7 and 8); N=33

Trials	2-4	5-7	8-10
5-7	.74		
8-10	.66	.88	
11-13	.64	.81	.84

TABLE 12 (*continued*)

INTERCORRELATIONS BETWEEN THE FORWARD ERRORS OF DIFFERENT GROUPS OF THREE TRIALS

Group I.4.Lb.A; N=41

Trials	2-4	5-7	8-10	11-13	14-16
5-7	.30				
8-10	.14	.49			
11-13	.04	.37	.42		
14-16	—.04	.24	.57	.37	
17-19	.06	.16	.42	.48	.50

Group I.4.Sc.B; N=31

Trials	2-4	5-7	8-10	11-13	14-16
5-7	.58				
8-10	.43	.48			
11-13	.07	.09	.21		
14-16	.02	.20	.41	.33	
17-19	.29	.35	.58	.18	.46

Group I.1.Sc.B'; N=31

Trials	2-4	5-7	8-10	11-13	14-16
5-7	.24				
8-10	.13	.46			
11-13	.46	.12	.48		
14-16	.33	.02	.38	.64	
17-19	.32	—.27	.30	.65	.64

Group II.1.Sc.D; N=32

Trials	2-4	5-7	8-10	11-13	14-16
5-7	.47				
8-10	.25	.42			
11-13	.37	.49	.61		
14-16	.00	.22	.36	.37	
17-19	.25	.19	.22	.05	.03

Group I.1.Sc.D

Trials	2-4	5-7
5-7	.09	
8-10	.21	.00

maze performance, however, has distinct merit in that it throws light on the question of what forces are operating to produce consistency between different groups of trials and between portions of the same series of trials. Thus, it can be seen from the tables that the highest coefficients are secured by the closely adjoining groups of trials and the lowest coefficients, in general, by the most remotely removed groups of trials. Hunter (7) called attention to the fact of this characteristic with his correlations of Vincent scores, and Tryon (29) has also noted the same characteristic with his data. Tryon has suggested that the proper interpretation of this characteristic of the data is that, apparently, different portions of the course of learning involve somewhat different functions, and seems to consider these high coefficients from closely adjoining trials to be fairly accurate measures of reliability. However, there is just as much evidence, or more, that this characteristic is to be regarded rather as evidence that various irrelevant factors (factors external to maze-learning ability, as such) operate to secure high correlations between closely related portions of the maze performance. These factors are more in the nature of systematic errors, e.g., position habits, raising the reliability coefficients, than they are of anything else.

In trying to utilize these correlations of groups of three trials as a basis for comparing the reliabilities found with different groups, the large number of coefficients involved makes it imperative to reduce the data to more manageable dimensions. Consequently, average values have been secured for the different

groups by the procedure of averaging the coefficients of alienation ($K=\sqrt{1-r^2}$) for the different groups. The advantage of this procedure over that of merely averaging the reliability coefficients is that reliability coefficients cannot be evaluated directly in terms of their relative size. A correlation of .80 is not twice as meaningful as one of .40, but more than twice as meaningful. The coefficients of alienation correct for this.

Table 13 shows the mean coefficients of alienation of the intercorrelations of each group of three trials with every other group of three trials, and the grand means for each group of rats and each group of trials. It is probably safe to say that we arrive by this means at a fairly satisfactory means of estimating the relative reliability of the different procedures used, insofar as the correlations from groups of three trials can indicate reliability. The groups of rats which yielded the highest reliabilities, judging from this table, are the 13-door groups, with the 1-door groups next, and the 4-door groups lowest. (The more reliable a test the lower its

TABLE 13

MEAN COEFFICIENTS OF ALIENATION OF ALL THE CORRELATIONS OF EACH GROUP OF THREE TRIALS WITH ALL OTHER GROUPS OF THREE TRIALS

Group	Groups of trials						Mean for
	2-4	5-7	8-10	11-13	14-16	17-19	all trials
I.13.*Sc.C*	.962	.848	.860	.844	.836	.847	.893
II.13.*Sc.C*	.926	.856	.799	.707	.755	.835	.813
I.4.*Lb.A*	.988	.943	.899	.928	.917	.928	.934
I.4.*Sc.B*	.948	.933	.898	.976	.945	.917	.936
I.1.*Sc.B'*	.948	.962	.926	.857	.881	.876	.908
II.1.*Sc.D*	.950	.924	.809	.904	.859	.984	.905
Mean for all groups	.954	.911	.865	.869	.865	.915	

coefficient of alienation.) As regards the different groups of trials, it seems that the groups 8-10, 11-13, and 14-16 seem the most reliable, but this apparent result may be the expression of the fact that these trials occur in the middle of the training series and are more linked by transfer effects and systematic errors to other groups of three trials than are the groups of three trials at either end of the training period.

In the course of this discussion we have suggested a number of times that poor experimental control could raise reliability coefficients in a manner that would not at all be justified. To demonstrate this, reliability coefficients have been secured by combining two of the groups which were run under the same maze conditions (four doors were used with each group), but with different feeding programs. Groups I.4.*Lb.A* and I.4.*Sc.B* were so combined, and reliability coefficients were calculated from the combined groups. It will be remembered that Group I.4.*Lb.A* had been fed rather liberally, and Group I.4.*Sc.B* rather scantily; hence the combined group gives much the same effect as would be secured if individual rats in the large group had been fed differently. The reliability coefficients from the combined group (see Table 14) are higher in every case than the correlation coefficients of either group separately, except for the correlation of trials 2-4 versus 5-7. This raising of the correlation is obviously the result of irrelevant factors existing with this combined group.

Still another influence which may seriously affect the accuracy of reliability coefficients is the extension

TABLE 14

CORRELATIONS OF ERRORS ON DIFFERENT GROUPS OF THREE TRIALS FOR GROUPS I.4.*Lb.A* AND I.4.*Sc.B* SEPARATELY, AND FOR THESE TWO GROUPS COMBINED

Trials correlated	Correlations I.4.*Lb.A*	I.4.*Sc.B*	I.4.*Lb.A* and I.4.*Sc.B*
2-4 *vs.* 5-7	.29	.58	.48
5-7 *vs.* 8-10	.49	.48	.64
8-10 *vs.* 11-13	.42	.21	.62
11-13 *vs.* 14-16	.37	.33	.55
14-16 *vs.* 17-19	.50	.46	.63

of the range of ability in a group by the inclusion of extreme cases. In *product-moment correlations* these affect the correlation in proportion to their extremeness. (More than one correlation has been cited throughout the previous pages in which a single case sometimes raised correlations as much as 20 or 30 points.) The product-moment correlation does not require normality of distribution of itself, and where these extreme cases are the *result of the same forces affecting the rest of the group, except that the forces are here operating* on a larger scale, such extreme cases should be allowed their unusual influence upon the correlations.

In maze experiments it is very doubtful whether such is the case. As an illustration, Group I.13.*Sc.E* may be used. From Table 15 it may be seen that the four extreme rats have had great weight in determining the correlations with this group, and that their omission leaves quite a different picture. It may also be seen that if rank-difference correlations were used throughout, a perhaps much more correct estimate of reliability would have been secured (with the total group) than

TABLE 15

GROUP I.13.*Sc.E*

Correlation coefficients calculated by the product-moment formula and by the rank-difference formula, with and without four extreme cases included. The upper coefficient in each pair is with the full group of 33 rats, and the lower coefficient with but 29 rats. The rank-difference coefficients are in parentheses.

Trials	2-4	5-7	8-10
5-7	.74 (.60) .53 (.41)		
8-10	.66 (.56) .34 (.35)	.88 (.54) .44 (.41)	
11-13	.64 (.39) .22 (.10)	.81 (.58) .26 (.33)	.84 (.42) .14 (.15)

if the product-moment correlations had been used. (It might be added here that rank-difference correlations have also been calculated for the odd *vs.* even trial correlations of Table 7, but that there Group I.13.*Sc.E* was the only group whose correlations were very definitely lower, and the relationship between the different groups was left as with the product-moment correlations.)

d. Correlations of scores on different mazes. The final set of reliability coefficients to present are those derived from tests and retests. Data are available on Group *C* and Group *D* (see Table 3). Group *D* was tested first for 30 trials on Maze II, and then, after 30 days of rest and 10 days of a second preliminary training, given 10 trials on Maze I, with but one door used in each maze. Group *C* was tested first on Maze I and then, after a similar interval, on Maze II, with 13 doors used with each maze. Group I.13.*Sc.E* was given 7 trials on Maze I, then 38 to 40 days of rest,

followed by 5 days of training on the straightaway, and then 6 more trials on the same maze. During the interval between the two periods of training the rats were fed an unlimited amount of dry food, and occasional bread, milk, and green food, *in order to bring the weights of all of them back to normal and thus reduce to a minimum the danger that differences in feeding would last over from the test to retest.* Restriction of feeding and weight drop began with the beginning of the second preliminary training in a manner similar to that used on the first preliminary training. Group *C* was also tested on the elevated maze.

The correlations were determined, in these cases, using those groups of trials which on various general grounds might have been thought to prove most enlightening. Thus, with time scores the first trial was omitted from the correlations because of the previous demonstration that its omission improved the reliability of time scores with most groups. The resulting correlations are given in Table 16.

It will be seen that there is quite a difference between Group *C* and Group *D* with respect to the correlations between the performances on Mazes I and II. In the case of Group *D* no one of the correlations is significant; in the case of Group *C* the correlations of scores on the two mazes, though appreciably lower than corresponding correlation coefficients from either the test or retest, are still fairly high. The intercorrelations of Maze I and Maze II with Group *D* similarly reflect the reliability coefficients from Maze I and Maze II separately, for the reliability coefficients for Group

TABLE 16

CORRELATIONS BETWEEN DIFFERENT MAZES, WITH INTERNAL CONSISTENCY CORRELATIONS FROM THE SAME GROUPS FOR COMPARISON

(See Table 12 for additional comparative data)

Trials Involved (except that the first trial is excluded from all time correlations)	Group I.13.*Sc.C* vs. II.13.*Sc.C* r's from time scores	I.13.*Sc.C* vs. II.13.*Sc.C* r's from error scores	II.1.*Sc.D* vs. I.1.*Sc.D* r's from time scores	II.1.*Sc.D* vs. I.1.*Sc.D* r's from error scores
Test-retest correlations:				
1-30 on I *vs.* 1-20 on II	.35±.15	.71±.08		
21-30 on I *vs.* 11-20 on II	.69±.09	.58±.11		
21-30 on I *vs.* 11-20 on II, but with one extreme case dropped	.34±.15			
11-20 on I *vs.* 11-20 on II	.26±.16	.56±.12		
Trials preceding severe norm on I *vs.* trials preceding severe norm on II		.30±.16		
Trials preceding moderate norm on I *vs.* trials preceding moderate norm on II		.36±.15		.28±.16
1-30 on II *vs.* 1-10 on I			—.02±.18	—.02±.18
1-20 on II *vs.* 1-10 on I				—.07±.18
1-10 on II *vs.* 1-10 on I				.04±.18
21-30 on II *vs.* 2-10 on I			—.08±.18	
Internal consistency correlations:				
1-30 on test, odd *vs.* even	.79±.07	.88±.04	.29±.16	.46±.14
Odd *vs.* even on retest (20 trials for II.13.*Sc.C* and 10 for I.1.*Sc.D*)	.76±.07	.88±.04	.21±.17	—.44±.14
1-10 *vs.* 11-20 (test)	.22±.16	.56±.12		
1-10 *vs.* 11-20 (retest)	.40±.14	.50±.13		
1-10 *vs.* 21-30 (test)	.17±.17	.30±.16		

D on the first maze learned were generally the lowest coefficients of all the groups, and on Maze I, on its retest, the internal-consistency measures were even lower (see Tables 7, 8, and 12). An analysis of the records of Group *D* on this second maze, however,

makes one suspect that the training on the first maze, with its reverse pattern, had a disturbing rather than a stabilizing effect on the learning of the second maze. This would seem the case because of the larger mean number of errors on Trial 1 on the second maze than on the first maze and the considerably greater variability of scores on this second maze on the first trial than on the first trial of the first maze. Hence the extremely low intercorrelations between the two mazes with Group *D* do not necessarily indicate that the reliability of the experimental data for Group *D* on the first maze is lower than one would have estimated from the reliability coefficients calculated on that maze alone.

The correlations between Group *C* on Maze II and on the elevated maze are negligible. The correlation between forward errors in Trials 1-30 on II and all entrances into blinds on the elevated maze was $.20 \pm .17$; and the correlation of the number of trials required to satisfy the norm of learning of three successive errorless runs, on the two mazes, was $-.07 \pm .18$. These results with the elevated maze do not necessarily indicate a low reliability for this type of maze.[9] It may be that what accounts for the low reliability indicated for this maze is in good part the procedure which was used of giving the trials one after another till learning was completed. Or the reason may have been that on the elevated maze the rats were run at more or less irregular times in the evenings of the day on which they had met the norm of learning on Maze II. After their

[9]See the article by W. R. Miles (20), for comments on some probable advantages of elevated mazes.

previous regular running in the afternoon, this may have caused some disturbance.

e. Correlations of test and retest. The correlations reported above of the same groups on Mazes I and II may be considered to be essentially test-retest coefficients, and to have all the advantages indicated in the early part of this paper for this method. However, the data from Group I.13.*Sc.E* is here presented separately because it conforms exactly to the test-retest formula. The internal-consistency coefficients from this group are gathered together in Table 16 for comparison with the test-retest coefficients. Figures are presented both for the entire group and for the group with the four extreme cases omitted. (It will be remembered that an interval of 43 to 45 days separates Trials 7 and 8.)

TABLE 17

TEST-RETEST COEFFICIENTS FROM GROUP I.13.*Sc.E*, WITH INTERNAL-CONSISTENCY CORRELATIONS FROM THE SAME GROUP FOR COMPARISON

	With the four extreme cases omitted		With the entire group	
Trials correlated	*r*'s from time scores	*r*'s from error scores	*r*'s from time scores	*r*'s from error scores
Test-retest:				
2-7 *vs.* 8-13	.61±.11	.53±.13	.74±.08	.83±.05
2-4 *vs.* 8-10		.34±.16		.66±.10
2-4 *vs.* 11-13		.22±.18		.64±.10
5-7 *vs.* 8-10		.44±.15		.88±.04
5-7 *vs.* 11-13		.26±.17		.81±.06
Internal-consistency correlations:				
Odd *vs.* even in 2-7	.29±.16	.57±.11	.81±.06	.78±.07
Odd *vs.* even in 8-13	.76±.07	.34±.16	.77±.07	.88±.04
2-4 *vs.* 5-7		.53±.13		.74±.08
8-10 *vs.* 11-13		.14±.18		.84±.05

It may be seen from the above table that the test-retest correlations compare rather favorably with the analagous internal-consistency correlations, just as the correlations of the scores of I.13.*Sc.C* with the scores of II.13.*Sc.C* compared favorably with the internal-consistency correlations. This result is rather in disagreement with the findings of Heron (6) and makes it seem probable that the intervals of time used by Heron were so long as to permit changes of relative vigor and health which would lower the correlations.

V

GENERAL SUMMARY AND CONCLUSIONS

The theoretical discussion of the present paper has led to the following conclusions:

1. To measure the reliability of a difference between group means in maze experiments, the formula which should be used is the customary one,

$$\sigma_{M_1 - M_2} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2},$$

or, where it is possible to correlate the scores of the two distributions, the formula,

$$\sigma_{M_1 - M_2} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r_{1.2}\,\sigma_{M_1}\sigma_{M_2}}.$$

The formula recently proposed by Tryon for the reliability of a difference,

$$\sigma_{M_1 - M_2} = \sqrt{\left(\frac{\sigma_1\sqrt{r_1}}{\sqrt{N_1}}\right)^2 + \left(\frac{\sigma_2\sqrt{r_2}}{\sqrt{N_2}}\right)^2},$$

is not the proper formula to use in this connection, because *it rests on the assumption that the standard deviation of found means around the true mean would be constant regardless of the reliability of the measuring instrument* and because this assumption is not in any sense correct.

2. The value of reliability coefficients relative to maze experiments is chiefly as a means of estimating the relative usefulness for experimental work of different maze patterns and procedures. The reliability

coefficients are not required as a means of estimating the reliability of a difference between group means, though certain other precautions are much more important than is at present generally recognized—particularly the necessity of securing randomness of sampling and of making the conditions strictly comparable between the compared groups.

3. When maze scores are as markedly skewed as is generally the case with current maze procedures, the median is a more reliable measure of central tendency than is the arithmetic mean.

4. The reliability coefficients from maze experiments cannot be interpreted as indicating directly the reliability of those experiments, because

a. The scores correlated to secure reliability coefficients in maze experiments do not satisfy the criterion of being independent measures of the same thing. Some methods of calculating maze reliability coefficients, however, more nearly approximate this ideal than do others, and accordingly are to be advised—especially the methods of correlating scores on test and retest, on different groups of trials within the test or retest, and possibly on odd and even trials. The reliability coefficients from correlation of scores on odd and even blinds, or on the first half *vs.* the second half of the maze, however, are to be recognized as seriously objectionable.

b. The size of reliability coefficients from maze experiments is generally significantly influenced by the degree of heterogeneity of ability of the group tested. Moreover, there is at the present time no statistical for-

mula which can be used to make possible the accurate comparison of reliability coefficients from groups of different degrees of heterogeneity. The formula commonly used for this purpose with educational and intelligence tests,

$$\frac{\sigma}{\Sigma}=\frac{\sqrt{1-R}}{\sqrt{1-r}},$$

is not applicable because of the fact that in maze experiments the standard errors of scores are not uniform from group to group, but vary greatly in response to the differences in the factors governing the height of the learning curve. Accordingly, a particularly important aspect of technique in experiments aiming to determine the relative reliability of different maze procedures is the equating of groups with regard to heterogeneity of ability.

5. The interpretation of the significance of maze scores awaits the empirical determination of the validities of the same.

In the experimental work, the object has been to determine the influence on the reliability of maze experiments of various features of maze structure and maze procedure. In particular the objects were (1) to determine the influence of the use of various numbers of doors to prevent retracing, (2) to determine the relation of methods of feeding to reliability, and (3) to secure more information on test-retest correlations with this maze. For these experiments the multiple-T maze of Stone and Nyswander was used, and their procedure was closely followed except that for preliminary train-

ing a short straightaway maze was used rather than a problem box, and that a special room around the maze was used to shelter the rats from distracting sounds during runs. Six groups of from 31 to 41 rats each were used.

The results of the experiment are presented not only on reliability, but also on the learning curves, variability of scores, and effects of differences in feeding, the reason for presenting all these data being the aid which this material gives in interpreting the reliability coefficients secured. It was found that, first, the learning curves after the first two trials were much lower than those of Stone and Nyswander's groups (see Figure 7). Comparison of procedures would seem to indicate that the difference in preliminary training may have accounted for this. Secondly, the variability of the scores was directly related to the height of the learning curves, being greater with the higher curves (see Figure 10 and Table 5 in comparison with Figure 7). Thirdly, after about the first seven trials the distribution of scores became more and more markedly skewed (see Figure 11). This skewness in itself indicates the need for longer and more difficult mazes to secure the maximum reliability. Fourthly, when the scores on the first trial were correlated with the scores on subsequent trials, only negligible correlations were found, except with two of the 1-door groups (see Table 6). With these two groups there were found correlations of $-.32 \pm .16$ and $-.42 \pm .15$ between forward errors on the first trial and forward errors on subsequent trials. These correlations indicate that where retracing is permitted, an

appreciable and variable amount of training may go on in the first trial, and that, consequently, the first trial, with such a maze, cannot be dropped as though it were roughly a constant. When we add to this observation the further considerations that the time and error scores on the first trial are largely determined by chance and that in a maze where retracing is permitted these scores on the first trial are so large and variable as to dominate the total scores, it can be clearly seen that there are serious objections to permitting retracing. Fifth, practically zero correlations were found between error scores and percentage loss of weight (see Table 7), demonstrating adequate control of feeding and motivation by the system used. By combining scores from a liberally fed and a scantily fed group, however, it was demonstrated that the reliability coefficients could have been unwarrantedly high if there had been a less careful control of feeding (see Table 14).

Reliability coefficients were calculated for both time and errors from odd *vs.* even trials, groups of 10 or 15 trials against one another, and (for errors) by the correlations of various groups of three trials within Trials 2-19. To secure test-retest data, two of the groups were run, after 30 days of rest and unrestricted feeding after training on a first maze, on a second maze which was a mirror image of the first, and a third group was given further training on the same maze after a first training of 7 trials and an interval of 38 to 40 days of rest and unrestricted feeding.

In general, the reliability coefficients from this study are definitely lower than those found with the same

maze pattern by Stone and Nyswander and by Heron. This may have been the result, to some extent, of the removal, especially by the careful control of feeding, of certain systematic errors which may have entered into their work to raise the correlations. A more probable explanation is that whatever forces produced the lower learning curves of the present study also caused the lower reliability coefficients, as these forces presumably would operate in the direction of making the problem essentially so much easier that it no longer had the same capacity for differentiating between the different animals.

In comparing the reliability coefficients from different groups to determine which particular variations of maze procedure and which type of maze scores afford the most reliable measures, it is to be noted that there are some contradictions between the reliability coefficients calculated by different methods. Also it has been necessary to interpret the correlations with caution because of the tendency in some cases for extreme scores to dominate the correlations. What I have sought, therefore, as the basis for my conclusions, has been the observation of consistency of trend, together with inspection of the scatter diagrams to determine which correlations seemed dominated by a few cases, and which seemed more the expression of a tendency in the entire group (and accordingly more dependable).

When the data are so considered (see Tables 8, 10, 12, 16, and 17 for reliability coefficients from error scores, and Tables 9, 11, 16, and 17 for reliability co-

efficients from time scores), the following conclusions are indicated:

1. With the multiple-T maze, prevention of retracing with 13 retracing doors gives higher reliabilities than when 4 doors are used, and this in turn gives definitely higher reliabilities than the use of a door only at the food box.

2. Strong motivation seems to yield more reliable results than moderate motivation.

3. Time scores are somewhat less reliable than error scores.

4. The reliability of time scores is increased if the first trial is not included in the calculations, except perhaps with groups where retracing is limited (see Table 9).

5. Judging from correlations of test and retest on different mazes, scores in terms of trials-to-learn are less reliable than error scores (see Table 16).

6. The correlation of different parts of the same period of training (whether by correlation of odd and even trials, or of groups of 3, 10, or 15 trials) tends to give higher reliability coefficients than are actually warranted by the reliability of the experiment. Three lines of evidence bear on this: (*a*) test-retest correlations on different mazes yielded lower figures than the internal correlations from either test or retest (see the figures for Group I.13.*Sc.C*, Table 16, as illustrating this, even though the internal correlations have not been corrected for halving of the data; (*b*) the size of the correlations of different groups of three trials with each other varied quite consistently and in an inverse man-

ner with the number of trials separating the trials correlated (see Table 12); and (*c*) in contrast with the above, correlations of test and retest on the *same* maze (with one group thus tested) yielded coefficients that compared well with the internal-consistency correlations of the same group (see Table 17).

This sixth conclusion is different from the position taken by Tryon and some other investigators, who would tend rather to accept the reliability coefficients from maze experiments at their face value. Such a procedure would be justified only if such coefficients were derived from scores that were really independent measures of the same thing—only if the scores correlated were not linked by "systematic errors" (i.e., by factors irrelevant to basic maze-learning ability—factors such as motivation differences, differences in emotional conditioning to handling or to the apparatus, etc.). This comment may raise some protest from those who are accustomed to say that an instrument is reliable if it measures reliably that which it does measure. Let us, therefore, examine the sense in which reliability coefficients are actually used. Thus, suppose that with an educational test there have been found high reliability coefficients from the correlation of odd *vs.* even items, but low correlations of test and retest. Suppose that it is then further discovered that this contrast results from the fact that the interpretation of instructions vitally affects the scores on the test and that this interpretation is liable to vary with the same person from time to time, and also that the interpretation of instructions varies from person to person. It does not

matter then that the test on each occasion "measures accurately that which it does measure," the test is condemned because it does not measure the same thing at different times or in different individuals. Exactly the same situation prevails in maze experiments, as we can conclude now not only from general theoretical arguments, but also from the general weight of the experimental data referred to above.

7. The reliability of a maze experiment is a function not merely of the maze pattern, but also of many different aspects of the maze procedure, such as the preliminary training, motivation, etc., even when these various aspects are carefully controlled. I derive this conclusion from the comparison of my reliability coefficients with those secured by Stone and Nyswander (24) and Heron (6). I have no reason to believe that my rat groups were more homogeneous than theirs or that my maze was easier to learn because of the 4" greater length of the alley units. The possibly significant differences which might account for the lower reliability coefficients of the present study must, therefore, be either the use of the sound-deadening room or the difference in preliminary training, or perhaps the securing of a much more nearly optimum feeding program than was used by Stone and Nyswander, despite the fact that they felt they had secured an optimum procedure. It has been suggested to me by Dr. Nyswander that the room may have served to lower the reliability coefficients by producing some degree of emotional disturbance rather than to eliminate distractions. Such an interpretation would not seem warranted, however,

because of the fact that my learning curves are so much lower than those of Stone and Nyswander. Also, I understand that the rats of Stone and Nyswander were run under quite quiet conditions. It may be, therefore, that merely the difference in preliminary training produced the difference in reliability coefficients.

Finally, attention should be called again to the very great desirability of using the split-litter technique wherever group comparisons are to be drawn. I base this conclusion upon the slight differences between the learning curves and reliability coefficients of Groups I.4.*Sc.B* and I.1.*Sc.B'* (which were run by the split-litter technique) as contrasted with the greater differences between other groups supposedly paired in conditions, such as Groups I.1.*Sc.B'* and II.1.*Sc.D*, and Groups I.13.*Sc.C* and I.13.*Sc.E*. This is the conclusion also reached by Corey (2) in his experiments on the effects of muscular activity on the rate of learning.

References

1. Carr, H. A. The reliability of the maze experiment. *J. Comp. Psychol.*, 1926, 6, 85-93.
2. Corey, S. M. Equating groups in comparative experiments. *J. Comp. Psychol.*, 1930, 10, 287-294.
3. Garrett, H. E. Statistics in psychology and education. New York: Longmans, Green, 1926. Pp. xiii+317.
4. Heron, W. T. The reliability of the inclined plane problem box as a measure of learning ability in the white rat. *Comp. Psychol. Monog.*, 1922, 1, No. 1, 1-36.
5. ———. Individual differences in ability vs. chance in the learning of the stylus chance. *Comp. Psychol. Monog.*, 1924, 2, No. 8. Pp. 60.
6. ———. The test-retest reliability of rat learning scores from the multiple-T maze. *J. Genet. Psychol.*, 1930, 38, 101-113.
7. Hunter, W. S. Correlation studies with the maze in rats and humans. *Comp. Psychol. Monog.*, 1922, 1, No. 1, 37-56.
8. ———. A reply to Professor Carr on "The reliability of the maze experiment." *J. Comp. Psychol.*, 1926, 6, 393-398.
9. Hunter, W. S., & Randolph, V. Further studies on the reliability of the maze with rats and humans. *J. Comp. Psychol.*, 1924, 4, 431-442.
10. ———. A note on the reliability of the maze as a method of learning in the Angora goat. *J. Genet. Psychol.*, 1926, 33, 3-8.
11. Husband, R. W. A comparison of human adults and white rats in maze learning. *J. Comp. Psychol.*, 1929, 9, 361-377.
12. Kelley, T. L. Statistical method. New York: Macmillan, 1923. Pp. v+390.
13. King, H. A. Growth and variability in the body weight of the rat. *Anat. Rec.*, 1915, 9, 751-776.
14. Lashley, K. S. A causal factor in the relation of the distribution of practice to the rate of learning. *J. Anim. Behav.*, 1917, 7, 139-142.
15. ———. Brain mechanisms and intelligence. Chicago: Univ. Chicago Press, 1929. Pp. vii+186.
16. Liggett, J. R. A note on the reliability of the chick's per-

formance in two simple mazes. *J. Genet. Psychol.*, 1925, 32, 470-480.

17. ———. An experimental study of the olfactory sensitivity of the white rat. *Genet. Psychol. Monog.*, 1928, 3, 1-64.
18. ———. The unit maze: a study in maze method. *J. Genet. Psychol.*, 1930, 37, 163-175.
19. MAURER, S. & TSAI, L. S. Vitamin B deficiency in nursing young rats and learning ability. *Science*, 1929, 70, 456-458.
20. MILES, W. R. The comparative learning of rats on elevated and alley mazes of the same pattern. *J. Comp. Psychol.*, 1930, 10, 237-261.
21. MURCHISON, C. [Ed.] The foundations of experimental psychology. Worcester, Mass.: Clark Univ. Press, 1929. Pp. x+907.
22. PATERSON, D. G. The Johns Hopkins circular maze studies. *Psychol. Bull.*, 1917, 14, 294-297.
23. SHEN, E. The standard error of certain estimated coefficients of correlation. *J. Educ. Psychol.*, 1924, 15, 462-465.
24. STONE, C. P. The reliability of rat learning scores obtained from a modified Carr maze. *J. Genet. Psychol.*, 1928, 35, 507-521.
25. STONE, C. P., & NYSWANDER, D. The reliability of rat learning scores from the multiple-T maze as determined by four different methods. *J. Genet. Psychol.*, 1927, 34, 497-524.
26. TOLMAN, E. C. The inheritance of maze-learning ability in rats. *J. Comp. Psychol.*, 1924, 4, 1-18.
27. TOLMAN, E. C., & DAVIS, F. C. A note on the correlations between two mazes. *J. Comp. Psychol.*, 1924, 4, 125-136.
28. TOLMAN, E. C., & NYSWANDER, D. The reliability and validity of maze-measures for rats. *J. Comp. Psychol.*, 1927, 7, 425-460.
29. TRYON, R. C. Effect of the unreliability of measurement on the difference between groups. *J. Comp. Psychol.*, 1926, 6, 449-453.
30. ———. Individual differences at successive stages of maze learning. Ph.D. Thesis deposited in the Library of the University of California, Dec. 1928. Pp. 113. (Most of the material of this thesis is summarized in 33*a*.)
31. ———. Demonstration of the effect of unreliability of measurement on a difference between groups. *J. Comp. Psychol.*, 1928, 8, 1-22

32. ———. Errors of sampling and of measurement as affecting difference between means. *J. Comp. Psychol.*, 1929, 9, 191-195.

33. ———. The genetics of learning ability in rats; preliminary report. *Univ. Calif. Publ. Psychol.*, 1929, 4, 71-89.

33*a*. ———. Studies in individual differences in maze ability. I. The measurement of the reliability of individual differences. *J. Comp. Psychol.*, 1930, 11, 145-170.

34. WALKER, H. M. Concerning the standard error of a difference. *J. Educ. Psychol.*, 1929, 20, 53-60.

35. YOSHIOKA, J. A preliminary study in discrimination of maze patterns by the rat. *Univ. Calif. Publ. Psychol.*, 1928, 4, 1-18.

36. ———. A further study in discrimination of maze patterns by the rat. *Univ. Calif. Publ. Psychol.*, 1929, 4, 135-153.

37. ———. Weber's law in the discrimination of maze distance by the white rate. *Univ. Calif. Publ. Psychol.*, 1929, 4, 155-184.

38. YULE, G. U. An introduction to the theory of statistics. (8th ed.) London: Griffin, 1927. Pp. xv+422.

LA CONSTANCE ET LA VALIDITÉ DES EXPÉRIENCES AVEC LE LABYRINTHE CHEZ LES RATS BLANCS

(Résumé)

Cette monographie contient (1) une analyse théorique des problèmes de la constance et de la validité comme les présente les expériences avec le labyrinthe et (2) un rapport d'une enquête expérimentale sur l'influence de certains facteurs sur la constance des expériences avec le labyrinthe T-multiple. Les sujets théoriques discutés sont: la signifiance de diverses mesures des différences entre les groups, surtout à l'égard de la formule récemment proposée par Tryon; la valeur de la médiane dans les données du labyrinthe en vue de l'écart de celles données; la valeur relative de différentes méthodes de calculer la constance des expériences avec le labyrinthe; et l'influence de la variation de la capacité sur les coefficients de constance obtenus.

Dans l'enquête expérimentale, l'appareil et la technique ont ressemblé à ceux de Stone et de Nyswander. Cependant, on a fait subir les épreuves préliminaires sur un parcours droit plutôt que sur une boîte d'échappement à plate-forme, et l'on s'est servi d'une chambre spécialement construite pour protéger le parcours droit et le labyrinthe contre les sons de l'extérieur. Le nombre des portes employées pour empêcher le retracement ont varié avec divers groupes; on a testé deux groupes avec une porte de retracement, deux groupes avec 4 portes de retracement, et trois groupes avec 13 portes.

Les courbes d'apprentissage sont beaucoup moins élevées que celles de Stone et de Nyswander. Avec la plupart des groupes les erreurs moyennes ont été moins élevées pour la dixième épreuve que les erreurs moyennes des groupes de Stone et de Nyswander à la treizième épreuve. Possiblement à cause de ceci, les coefficients de constance trouvés ont été un peu moins élevés que les leurs.

La variabilité des résultats a varié approximativement en proportion des erreurs moyennes, les groupes avec les courbes d'apprentissage les moins élevées ayant la plus petite variabilité. Dans toutes les épreuves sauf les premières les résultats sont écartés vers les plus grandes valeurs, l'écart devenant plus grand avec la continuation de l'entraînement. On montre que cet écart indique le besoin d'un labyrinthe plus long et plus complexe pour obtenir les constances maxima.

Les corrélations du pourcentage de la réduction du poids avec les erreurs sont si peu élevées qu'elles indiquent que le contrôle de l'alimentation a été suffisamment exact.

Les corrélations des erreurs de la première épreuve avec les erreurs des épreuves subséquentes, chez deux des groupes employant une porte, indiquent que les résultats de la première épreuve ne peuvent être négligés où le retracement n'est pas empêché, comme si leur influence sur l'apprentissage était constante.

La corrélation des erreurs dans différents groupes d'épreuves du même labyrinthe indique que les constances les plus élevées sont obtenues par le labyrinthe à 13 portes et les constances les moins élevées par le labyrinthe à une porte. Chez les groupes courant avec divers degrés de réduction du poids, une motivation forte paraît plus favorable à la constance qu'une motivation faible.

Les corrélations entre les erreur de deux labyrinthes T-multiple avec un des groupes employant 13 portes indiquent qu'on peut obtenir une constance

assez satisfaisante avec le labyrinth T-multiple. Les essais d'apprentissage paraissent un type de résultat moins constant que les erreurs totalles.

LEEPER

ZUVERLÄSSIGKEIT UND GÜLTIGKEIT BEI LABYRINTHUNTERSUCHUNGEN (MAZE EXPERIMENTS) AN WEISSEN RATTEN

(Referat)

Diese Monographie enthält (1) eine theoretische Analyse der Probleme der Zuverlässigkeit und der Gültigkeit (validity), wie sie sich in Labyrinthuntersuchungen darstellen, und (2) einen Bericht über eine experimentelle Untersuchung des Einflusses gewisser Gegenstände auf di Zuverlässigkeit bei Experimenten mit dem multiplen T-formigen Labyrinth (multiple T maze). Die besprochenen theoretischen Fragen sind: die Bedeutung verschiedener Massstabe der Gruppenunterschiede, besonders mit Bezug auf die neulich von Tryon vorgeschlagene Formel; der Wert des Mittelwertes (median) bei Labyrinthbefunden, unter Edwägung der verzerrten Häufung (skewness) solcher (statistischer) Befunde; der relative Wert der verschiedenen Methoden zur Ermittelung der Zuverlässigkeit von Labyrinthexperimenten; und, schliesslich, die Einwirkung der Erstreckung (range) der Fähigkeitszahlen (ability scores) auf die ermittelten Zuverlässigkeitskoeffizienten.

Das bei der experimentellen Untersuchung gebrauchte Apparat und Verfahren gleichte dem von Stone und Nyswander gebrauchten. Die Vorversuche wurden aber nicht auf einer Schachtel mit Fluchplattform (platform escape box), sondern auf einer geraden Strecke (straightaway) gemacht, und man gebrauchte eine besonders gebaute Kammer (room) um die gerade Strecke und das Labyrinth von äusseren Geräuschen zu schützen. Die Zahl der zur Verhütung des Zurückgehens (retracing) benutzten Türen war bei den verschiedenen Gruppen verschieden. Bei der Prüfung von zwei Gruppen wurde nur eine einzige Wiederholungstür (retracing door) gebraucht, bei zwei weiteren Gruppen wurden vier Wiederholungstüren, und drei Gruppen dreizehn Türen gebraucht.

Die Lernkurven sind bedeutend niedriger als die von Stone und Nyswander erhaltenen. Bei den meisten Gruppen war zahl die der mittleren Fehler am zehnten Versuch schon niedriger als bei den Versuchen von Stone und Nyswander am dreissigsten Versuch. Vielleicht deshalb waren die erhaltenen Zuverlässigkeitskoeffizienten etwas niedriger als die ihrigen. Die Veränderlichkeit (variability) der Zahlen (scores) verhielt sich im Allgemeinen proportional zu den mittleren Fehlerzahlen, wobei die Gruppen mit niedriger-stehenden Lernkurven die geringere Veränderlichkeit erwiesen. In allen Versuchen mit Ausnahme der allerersten zeigen die Zahlen eine Verschiebung (skewness) nach der Richtung der höheren Werte hin—und diese Verschiebung wurde im Lauf der Dressierung ausgesprochener. Es wird darauf hingewiesen, dass diese Verschiebung die Nötigkeit eines längeren und viel komplizierteren Labyrinthes zur Erreichung maximaler Zuverlässigkeitszahlen andeutet.

Die Niedrigkeit der Korrelationen zwischen dem Prozentsatz des Gewichtverlustes und den Fehlern weisen darauf hin, dass die Kontrolle der Fütterung genügend genau war.

Korrelationen zwischen den Fehlern bei dem ersten Versuch und den Fehlern an wieteren Versuchen an zwei der ein-türigen Gruppen weisen

in, dass wo die Wiederholung nicht vermieden wird, die Resultate n Versuches nicht abgelegt werden können als ob ihre Einwirkung Lernen eine konstante sei.

orrelation unter den Fehlern bei verschiedenen Gruppen der Ver- dem selben Labyrinth weist darauf hin, dass die höchsten Zuver- szahlen bei der 13-Türigen Anordnung und die niedrigsten igkeiten bei der ein-Türigen Anordnung erzielt werden.

sultate bei Gruppen, in denen verschiedene Grade des Gewichts- bestanden, scheinen anzudeuten, dass starker Antrieb (motivation) lässigkeit eher begünstigt als schwacher.

ationen zwischen den Fehlern an zwei multiplen T-formigen Laby- ei eine der 13-Türigen Gruppen weisen darauf hin, dass mit dem T-formigen Labyrinth eine ziemlich befriedigende Zuverlässigkeit werden kann. Als Basis der Berechnung scheint die Zahl der zum ötigen Versuche weniger zuverlässig zu sein als die Gesamtzahl der

LEEPER

Journal of General Psychology

EDITED BY
CARL MURCHISON, *Clark University*

Quarterly, published in January, April, July, and October. Devoted primarily to experimental, theoretical, clinical, and historical psychology. Manuscripts may be sent to any member of the Editorial Board, or may be sent directly to the general editorial office. All subscriptions and business communications should be sent directly to the Clark University Press. Beginning January, 1927, one thousand pages (two volumes) annually. Per annum $14.00; per volume $7.00; single numbers $4.00. Complete sets from 1927 at $7.00 and $8.00 per volume, plus transportation.

The Pedagogical Seminary and Journal of Genetic Psychology

EDITED BY
CARL MURCHISON, *Clark University*

Quarterly, published in March, June, September, and December. Devoted to child behavior, animal behavior, and comparative psychology. Manuscripts may be sent to any member of the Editorial Board, or may be sent directly to the general editorial office. All subscriptions and business communications should be sent directly to the Clark University Press, [illegible]. Beginning [illegible], one thousand pages (two volumes) annually. Per annum $14.00; per volume $7.00; single numbers $4.00. Complete sets from 1891 at $7.00 and $8.00 per volume, plus transportation.

$7.00 per volume
Single numbers $2.00

MONTHLY
Two volumes per year

April, 1932
Volume XI, No. 4

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

A CRITICAL STUDY OF TWO LISTS OF BEST BOOKS FOR CHILDREN*

Department of Education, Yale University

By

FRANK K. SHUTTLEWORTH

*Recommended by Mark A. May, accepted by Carl Murchison of the Editorial Board, and received in the Editorial Office, February 13, 1931.

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the postoffice at Worcester, Mass., under Act of March 3, 1879

ACKNOWLEDGMENTS

In the prosecution of these studies, the author acknowledges his indebtedness to Professor Edwin D. Starbuck, formerly Director of the Institute of Character Research at the State University of Iowa, now at the University of Southern California. Professor Starbuck provided the opportunity and supplied the necessary facilities with a generous hand.

Professor Karl Holzinger of the University of Chicago read the material of Chapters II, III, and IV and made helpful suggestions.

Miss Ivy C. Husband and Mr. Donald Kiesau assisted materially in relieving the author of much of the detailed tabulations involved.

A special acknowledgment of indebtedness is due to the compilers of the *Winnetka Graded Book List*, Carleton Washburne and Mabel Vogel. They spared no pains in answering questions, retabulating material, and making available all of their original data. Their cooperation and generosity establish easily the high-water mark for such behavior over anything which the author had experienced or even hoped for.

FRANK K. SHUTTLEWORTH

YALE UNIVERSITY
NEW HAVEN, CONNECTICUT

CONTENTS

	Acknowledgments	249
I.	Introduction	253
II.	Agreements and Disagreements as to Grading of Books	257
III.	Explanations of Discrepancies in Absolute Grading: The Distribution of Ballots	264
IV.	Explanation of Discrepancies: Mode versus Median	273
V.	Judging the Relative Worth of Books	283
VI.	Agreements between Merit as Estimated by Competent Opinion and by Interest Value	290
VII.	Experimental Results: The Correlation between Judgments of Merit and Interest Value	298
VIII.	Experimental Results: Grade Placement and Interest Value	304
IX.	Summary, Interpretations, and Problems	312
	References	317
	Résumé en français	318
	Referat auf deutsch	319

I

INTRODUCTION

This monograph presents a comparative and critical study of two of the most elaborate of recent attempts in the compilation of lists of recommended books for children: The *Winnetka Graded Book List* (8) and *A Guide to Literature for Character Training* (4, 5). Each purports to be the last word in the application of scientific methods, yet each travels its own sweet way in producing its list. Do the lists agree? If not, why not?

Two undertakings more opposed in basic assumptions could hardly be discovered. The *Winnetka Graded Book List* is founded upon the child, the books he reads, and likes or dislikes, and what he thinks of them. Only incidentally and in negative fashion do adults tamper with what the child likes. The volumes of *A Guide to Literature for Character Training* are founded on adult opinion, on what the adult thinks is good for or interesting to the child, and are offered to adults as a finding-list of best books. Not even in a negative or supplementary fashion has the child a direct voice in determining what is recommended for him.

The contrasting assumptions of the *Winnetka Graded Book List* and *A Guide to Literature for Character Training* are but another chapter in the familiar conflict of educational philosophies. Shall the school be child-centered or adult-centered? Shall the curricu-

lum serve the child's immediate needs or shall it prepare the child for adult needs? Shall adults prepare the mental food of children according to adults' tastes or according to children's tastes? The purpose of this study is to inquire in what respects these contrasting assumptions give the same or different results.

The fundamental differences in the procedures of the two undertakings need to be clearly understood. The significant points in the Winnetka study are the following:

1. The primary data consist of over 100,000 ballots from 36,000 children. Each child filled out a ballot for every book read during the school year. The books, some 800, which received 25 or more ballots constituted the preliminary selection of best books.

2. Thirteen children's librarians were given this list of 800 books and asked to judge their literary merit and general worth. One hundred ten books which three-fourths of the librarians rated as either "not recommended because of low literary merit" or as "not recommended because of subject matter" were excluded.

3. The remaining 686 books are listed in order according to a popularity index which consisted of the product of the number of cities in which the book was read multiplied by the number of children who read and liked it. Supplementary to this order of listing, the books which were approved by three-fourths of the librarians were starred.

4. All children were given the paragraph-meaning section of the Stanford Silent-Reading Test. The re-

sulting scores were translated into values indicating the equivalent reading grade ability in terms of school grades. The median of the reading grade ability of all children reading and liking a book determined the school grade in which it was placed.

By these procedures the basic list of 800 books was determined by the number of children reading each book, the listing of the books according to a popularity index was determined by children, and the grading of each book was determined by the median reading grade ability of the children reading it. Only in elimination of 110 books did competent adult opinion enter into the picture.

When the important features of the volumes of *A Guide to Literature for Character Training* are considered, competent opinion determines every step.

1. A specific field is defined, i.e., fairy tale, or fiction, or biography; all available lists and catalogues are canvassed for suitable titles; and, as far as possible, the actual books are assembled in the Institute of Character Research at the State University of Iowa.

2. Each book is carefully read by at least three readers of the staff of the Institute and detailed judgments as to its literary merit and character value, its most suitable grade, ethical situation, moral attitudes, etc., are recorded.

3. About one-half of the books subjected to this detailed analysis are recommended according to five levels of merit, the non-recommended materials falling into four levels of demerit.

The data herein reported fall naturally into three

parts. Chapters II, III, and IV considered agreements as to placement of books in the several school grades. Chapters V and VI consider agreements as to recommendations of merit. Chapters VII and VIII report two supplementary studies designed to yield more precise answers to the major questions which are involved.

II

AGREEMENTS AND DISAGREEMENTS AS TO GRADING OF BOOKS

Do the two lists of recommended reading materials agree as to the grading of their books? If not, why not?

Accurate placing of children's books in the several grades involves two quite distinct problems. One problem is that of relative grading. Here the concern is to test, for example, whether five books, *A, B, C, D,* and *E* should be read in this order from the earlier to the later grades or in some other order. Correlations between the Winnetka list and the Guide should determine whether they agree as to relative grading. When the order *A, B, C, D,* and *E* is established there arises a second and very different problem, that of absolute grading. Should these books go into Grades II, III, IV, V, and VI, or into Grades V, VII, VIII, and IX? Here the question is whether the lists agree as to the average grade placement of their materials. Should they go into Grades II, IV, VI, VIII, and X, or into Grades IV, V, VI, VII, and VIII? Here the question is whether the lists agree as to the range of grades employed.

The *Winnetka Graded Book List* and *A Guide to Literature for Character Training* agree reasonably well as to relative grade placement. One hundred fifty-nine books are common to the Winnetka study and

Volume I of the *Guide*.[1] The correlation between the grading of the two is .665. Since the correlation is influenced by the range of grades involved, a more accurate statement is given by the probable error of estimate. Predicted from the Winnetka list, this is .82 of a grade. That is, if a mathematically best prediction of the *Guide* were made from the Winnetka list, 50% of the books would show differences of less than .82 of a grade. One hundred seventy-six books are common to the Winnetka list and Volume II of the *Guide*.[2] Here the correlation is .701 and the probable error estimate .84 of a grade. Combining the 159 and 176 cases gives a correlation of .812 with a probable error of estimate of .83 of a grade.

It is of some interest to know whether these figures hold true for comparisons of the *Guide* and of the Winnetka list with other lists based on competent opinion. For Volume I of the *Guide*, 346 books were found graded in one or more of 29 graded lists. Correlating the grading of the *Guide* against the average grading of these lists gave a correlation of .970. Table 1 displays an abbreviation of the scatter-diagram for the purpose of illustrating what is involved in well-nigh perfect agreement. Here the probable error of estimate of the *Guide* is .35 of a grade. These figures indicate that a combination of other graded lists must correlate with the Winnetka list about as does the *Guide*. The Winnetka grading has been checked

[1]Actually four lists are involved, since 110 books rejected from the Winnetka list and 399 which failed of recommendation in the *Guide* are employed.

[2]Only recommended books are employed here.

TABLE 1

SCATTERDIAGRAM SHOWING CORRELATION BETWEEN GRADING OF VOLUME I OF *Guide* AND TWENTY-NINE GRADED LISTS

A correlation of .970 was calculated from finer grouping. Italicized figures indicate perfect agreement.

Grading of Volume I of *Guide*	Twenty-nine published lists I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
XII												*3*
XI											*4*	
X									4	*5*		
IX								1	*3*	1		
VIII							5	*18*				
VII						21	*20*	2	1			
VI				1	7	*19*	1	1				
V				8	*68*	6	1	1				
IV			5	*65*	8	2						
III		5	*36*	4	1							
II	2	*10*	1									
I	*14*	2										

against three other graded lists. For 366 books common to the Winnetka list and the Wisconsin list (6), the correlation is .827 and the probable error of estimate from the Winnetka .88 of a grade. One hundred ninety-three books were found in both the Winnetka and Pittsburg (1) lists. Here the correlation is .820, the probable error of estimate from the Winnetka being .62 of a grade. Two hundred sixty-six books were found in the Winnetka and N.E.A. (2) lists. Here the correlation is .830, the probable error of estimate from the Winnetka being .75 of a grade. It is concluded from these correlations that the *relative* grading of the *Guide* and of the Winnetka list agrees reasonably well with other graded lists based on competent opinion.

Relative grading, however, is only one aspect of the

situation. The same correlations would be obtained if all the books in the Winnetka list or in the lists prepared by competent opinion were systematically lowered or raised any number of grades. The discrepancies appear in absolute grading. Here, for example, is the distribution of 159 books common to the Winnetka list and Volume I of the *Guide.*

Grades	I	II	III	IV	V	VI	VII	VIII	IX	Mean*	*S.D.*
Winnetka	0	0	3	37	72	44	2	1	0	5.57	.815
Volume I	8	16	24	41	43	14	9	2	2	4.73	1.622

*Means calculated assuming that books in Grade III, for example, cover the range from Grade 3.00 to Grade 3.99.

The important facts to note are the following. The average grade according to the Winnetka list is 5.57 while according to the *Guide* the average is 4.73, or almost a grade lower. The standard deviation, a measure of the range of grades, is .815 according to the Winnetka list and according to the *Guide* 1.622, or twice as much. Of the 159 books, 153 are concentrated in Grades IV, V, and VI by the Winnetka, while the *Guide* locates only 98 in these grades. Essentially the same contrast is given by the 176 books common to the Winnetka list and Volume II of the *Guide*:

Grades:	II	III	IV	V	VI	VII	VIII	IX	X	Mean	*S.D.*
Winnetka	0	0	2	12	24	78	47	13	0	7.61	1.031
Volume II	1	7	9	24	32	44	31	17	11	7.26	1.743

The discrepancies are not so marked but are of the same type. The Winnetka average is higher and the standard deviation much smaller. Of the 176 books,

149 are placed in Grades VI, VII, and VIII by the Winnetka, while the *Guide* places only 107 in these grades.

Comparing the Winnetka with the Wisconsin, Pittsburgh, and N.E.A. lists,[8] we have the following figures showing precisely the same tendencies:

Grades:	I	II	III	IV	V	VI	VII	VIII	IX	X	Total	Mean	*S.D.*
Winnetka:	0	0	2	39	72	56	58	32	9	0	268	6.47	1.39
N.E.A.	9	19	27	30	58	39	41	36	9	0	268	5.83	2.02
Winnetka	0	0	2	21	42	45	51	25	7	0	193	6.67	1.35
Pittsburgh	2	8	18	30	45	42	33	15	0	0	193	5.78	1.60
Winnetka	0	0	8	40	90	72	80	59	17	0	366	6.55	1.46
Wisconsin	29	15	24	46	67	44	52	37	40	12	366	6.11	2.41

A general tendency appearing in these data is that books which competent opinion places in Grades I, II, and III are located one or two grades higher by the Winnetka, while books which competent opinion places in Grades IX and X tend to be located a grade lower by the Winnetka. Nor does this tell the whole story, since the lists prepared by competent opinion on which we have reported are for the elementary grades. If lists for the high school are added, it appears that books placed in Grades IX, X, XI, and XII by competent opinion are on the average located in Grades VII and VIII by the Winnetka.

Table 2 in abbreviated form (compare with Table 1) gives the combined data showing agreements as to relative and absolute grading between the Winnetka

---

[8]In preparing the Wisconsin, Pittsburgh, and N.E.A. distribution, a book recommended for Grades III to V for example, is called a fourth-grade book, etc.

TABLE 2

SCATTERDIAGRAM SHOWING AGREEMENT BETWEEN THE WINNETKA LIST AND THREE LISTS REPRESENTING COMPETENT OPINION

Italicized figures indicate perfect agreement.

		Competent opinion									
		I	II	III	IV	V	VI	VII	VIII	IX	X
Winnetka	X										
	IX				1				8	*10*	11
	VIII					2	8	35	*32*	22	8
	VII			1	2	24	37	*48*	21	24	3
	VI			3	18	61	*31*	16	8	3	
	V	10	8	33	63	*42*	11	5	1		1
	IV	20	20	18	*12*	5	3	1			
	III	*7*	4								
	II										
	I										

and the two volumes of the *Guide* and the Wisconsin list. The Wisconsin list is selected because it is more recent and a larger number of books are involved. In all, 175 of the 701 books (some of which are duplicated) show perfect agreement, 312 show a difference of one grade, 158 show a difference of two grades, 43 books are displaced three grades, eleven are displaced five grades. The average difference is 1.16 grades. This is 47% reduction from a purely chance arrangement.

What must happen in order to adjust the absolute grading of the Winnetka to conform with competent opinion? The fourth-, fifth-, and sixth-grade books must be dropped one whole grade and the third-grade books dropped two whole grades. When this is done, the average discrepancy falls to .97 of a grade or a 59% reduction from purely chance arrangement. What must happen to adjust the grading of competent opinion

ɔnform to the Winnetka? The first-, second-, and
d-grade books must be moved up to Grade IV, and
ninth- and tenth-grade books placed in Grade VIII.
e the average difference falls to .61 of a grade or a
, improvement over chance arrangement. In the
:ess of altering the absolute grading, we have
:hed neither the relative gradings of the lists nor
r relative agreements. The gains in agreements are
entirely to changing the absolute grading. So far
ıpetent opinion and data from children do not give
same results. How are the discrepancies to be ex-
ned? This is the problem of the next two chapters.

III

EXPLANATION OF DISCREPANCIES IN ABSOLUTE GRADING: THE DISTRIBUTION OF BALLOTS

In the preceding chapter, evidence was presented showing that the relative grading of the *Winnetka Graded Book List* agrees reasonably well with *A Guide to Literature for Character Training* and that both agree with other lists based on competent opinion. Large disagreements, however, are found in absolute grading. In explanation of these discrepancies, the following pages point out two disturbing factors in the Winnetka study, one operating in the collection of the original data and another in its statistical treatment. Corrections for these factors have been confined to the 159 books common to the Winnetka list and Volume I of the *Guide* and their rejected lists. The corrected Winnetka grades which result are in almost perfect agreement with the grading of the *Guide*.

We shall consider in this chapter the failure of the Winnetka list to obtain an adequate sampling of ballots from children in the lower and higher grades. In searching for possible explanations of the large discrepancies between the grading of the *Guide* and the Winnetka list, the data of Table 3 seemed important. Table 3 quotes a statement and a table from page 14 of the introduction to the *Winnetka Graded Book List*.[4]

[4]Original edition.

It purports to display the number of ballots received from each grade of reading ability. According to the

TABLE 3

TABLE AND ACCOMPANYING STATEMENT QUOTED FROM PAGE 14 OF THE *Winnetka Graded Book List* PURPORTING TO GIVE THE NUMBER OF BALLOTS RECEIVED FROM CHILDREN OF EACH DEGREE OF READING ABILITY

Number of ballots. About 100,000 filled-out ballots were returned to us. Half of these (53,228 to be exact) were on 796 books, on each of which there were at least 25 ballots. The other half were scattered over 8,500 books, none of which had as many as 25 ballots. Of the 53,228 ballots on books read by 25 or more children, 22,184 were from boys and 31,044 from girls. These were distributed according to the reading grade of the children as follows:

TABLE NO. 1

	Boys	Girls	Totals
3rd	243	322	565
4th	2,972	4,186	7,157
5th	6,106	8,267	14,283
6th	3,775	6,867	10,642
7th	5,629	7,359	12,988
8th	2,978	3,509	6,487
9th	446	390	836
10th	125	145	270
	22,184	31,044	53,228

quoted table, no ballots were received from second-graders, only 565 ballots were received from third-graders, while fourteen, ten, and thirteen thousand ballots were received from fifth-, sixth-, and seventh-graders. These data seemed to explain the whole situation. Obviously, if no ballots were received from second-graders and only a very few from third-graders, there was no chance of any book being placed in the second grade and very little chance of any book being placed in the third grade.

This explanation, however, had to be discarded. Up-

on inquiry, Mr. Washburne wrote that the explanation of the table had been misstated. The table gave the number of ballots on books finally classified in the given grades. That is, there were 565 ballots returned on books finally classified in the third grade, and so forth.

The request was then made for the distribution by grades of the ballots. Mr. Washburne graciously consented to make the tabulation. The data given in Table 4 with the following explanation were received from Miss Vogel. The italics are hers.

> "The enclosed sheet entitled *Number of ballots on graded books* (see Table 4) gives the total number of ballots received on all books included in the graded book list and the excluded list. Read the table as follows: There were 1040 ballots received from children with second grade reading ability. This is 1.9% of the total number of ballots received (54,791). These 1040 ballots were on 256 different books, thus making the number of ballots received *per book* by second grade children, 4.1. Similarly, the the number of ballots *per book* filled out by third grade children was 6.2, by fourth grade children, 12.2, by fifth grade children, 13.2, etc. Because of the larger number of books reported on by children in the middle grades this measure is, of course, a much better one than the total number of ballots. The last columns entitled *Per cent of ballots per book* gives the proportion of *ballots per book* received from children of each reading grade.
>
> "Mr. Washburne and I talked the matter over with Professor - - - - of the University of Chicago. He advised us as follows: 'It seems to me that there should be a correction made for the varying percentage of ballots *per book*. You could do it on the books for third and fourth grades and see what difference it makes. I suspect the grade changes will be few. I should take *3* as a multiplier for ballots in grade two, and 2 as the factor for grade three, *leaving the rest up to grade nine unchanged. Thus*, if you found on a book 100 ballots for grade two, 400

TABLE 4

NUMBER OF BALLOTS, NUMBER OF BOOKS, AND NUMBER OF BALLOTS PER BOOK ACCORDING TO SCHOOL GRADE FOR 789 CASES OF BOOKS RECEIVING TWENTY-FIVE OR MORE BALLOTS

Reading grade	No. of ballots	% of ballots	No. of books	No. of ballots per book	% of ballots per book
II	1040	1.9	256	4.1	4.8
III	3043	5.6	491	6.2	7.4
IV	8491	15.5	699	12.2	14.5
V	10044	18.3	762	13.2	15.7
VI	9103	16.6	766	11.9	14.1
VII	8837	16.1	733	12.1	14.4
VIII	7225	13.2	687	10.5	12.5
IX	4096	7.5	576	7.1	8.3
X	1961	3.6	467	4.2	5.0
XI	951	1.7	389	2.8	3.3
Total	54,791				

ballots for grade three, and 100 for grade four, the change would give the following distribution: 300 for grade two, 800 for grade three, and 100 for grade four. This book would still be in grade three.' "

Table 4 shows very clearly that a relatively large number of ballots were obtained from Grades IV to VIII and a relatively small number from Grades II, III, IX, X and XI. The effect of this uneven distribution of ballots is to pull the preferred grade of all books toward the middle grades. What correction shall be applied to eliminate this factor?

The above letter suggests that the correction be made on the basis of the varying percentages of ballots per book. Specifically, it recommends that the second-grade ballots be multiplied by factor *3*, and the third-grade ballots by factor *2*, and presumably that the ninth-, tenth-, and eleventh-grade ballots be multiplied by factors of approximately *2*, *3*, and *4.5*. With the basis for

this correction the author disagrees. The specific corrections are not so seriously in error. They go in the right direction, but do not go far enough. The proper correction should be based on the percentage of ballots from each grade of reading ability. That is, instead of multiplying the number of ballots by the following factors based on the percentage of ballots per book as suggested:

Grades:	II	III	IV	V	VI	VII	VIII	IX	X	XI
Factors	3.2	2.1	1.1	1.0	1.1	1.1	1.3	1.9	3.1	4.8

the proper correction should employ the following factors:

Grades:	II	III	IV	V	VI	VII	VIII	IX	X	XI
Factors:	9.7	3.3	1.2	1.0	1.1	1.1	1.4	2.5	5.1	10.5

For Grades II and XI the proper correction is two or three times as severe.

The suggested correction was adopted upon an inadequate analysis of the situation. Evidently the reasoning employed was as follows. If the second-grade ballots are multiplied by factor *9.7* then the number of ballots per book read by second-graders would be 39.6, a number all out of proportion to the number of ballots per book read by fifth-, sixth-, and seventh-graders, where the similar numbers are 13.2, 11.9, and 12.1. The analysis should have proceeded a step further. If *only* books receiving at least 25 ballots originally are considered and factor *9.7* applied, then the minimum number of all the ballots for books read by second-graders would be *33.7,* while the minimum number of all the ballots on books read by fifth- and sixth-graders would continue to be 25. Obviously, un-

der these conditions the average number of ballots per book from second-graders must be much higher than for fifth- and sixth-grade children.

The assumption that the number of ballots per book *must be the same for all the grades holds, if it holds at all, only when we consider multiplying* *all* ballots by corrective factors. If this were done, then several hundred books which originally did not achieve 25 ballots would be candidates for inclusion in the list and the number of ballots per book might prove to be similar from grade to grade.

There is, however, a much more simple and direct way of evaluating the merits of the alternative corrections. So far the question has been, "What would have happened if 10,044 ballots had been received from each grade on books achieving at least 25 ballots?" If this question is restated, "What would have happened if only 951 ballots had been received from each grade on books achieving at least 25 ballots?" then the logic of the situation is no longer so involved. The proper correction would then be to *divide* by the following factors:

Grades:	II	III	IV	V	VI	VII	VIII	IX	X	XI
Factors:	1.1	3.2	8.9	10.6	9.6	9.3	7.6	4.3	2.1	1.0

If these factors are applied to the distribution of ballots for a given book and the average or median or modal grade determined, the result will be precisely the same as that obtained by multiplying by the more severe factors given on page 268.

There is a difficulty, however, in using the data of Table 4 since it gives the distribution of ballots only on

TABLE 5

NUMBER OF BALLOTS ACCORDING TO READING GRADE FOR ALL BOOKS BALLOTED UPON

Reading grade	No. of ballots
II	2010
III	6323
IV	14881
V	17584
VI	16452
VII	16237
VIII	13815
IX	8436
X	4571
XI	2431
Total	102741

books achieving 25 ballots. Accordingly, the request went forth for all the 50,000 ballots on books receiving less than 25 ballots and again in the finest spirit of co-operation the data were made available. With these ballots at hand, the data on a sample of every twelfth book were tabulated. Multiplying the results by 12 and combining them with the data of Table 4 gives the distribution of all ballots according to degrees of reading ability. These are displayed in Table 5. This provides the data necessary for an improved correction. The exact factors are as follows:

Grades:	II	III	IV	V	VI	VII	VIII	IX	X	XI
Factors:	8.8	2.8	1.2	1.0	1.1	1.1	1.3	2.1	3.8	7.2

In applying these corrections, in order to avoid cumbersome decimals, the actual factors used were:

Grades:	II	III	IV	V	VI	VII	VIII	IX	X	XI
Factors:	8.5	2.5	1.0	1.0	1.0	1.0	1.0	2.0	3.5	7.0

Application of these corrections to the distribution of ballots on two books will illustrate their effect.

TABLE 6

ORIGINAL DISTRIBUTIONS AND CORRECTED DISTRIBUTIONS FOR TWO BOOKS, ILLUSTRATING THE EFFECT OF CORRECTING FOR THE ORIGINAL INADEQUATE SAMPLING FROM THE DIFFERENT GRADES

Bow-wow and Mew-Mew—Craik

Grade:	II	III	IV	V	VI	Median grade
Original distribution	10	15	28	14	2	4.36
Effect of correction	85	37	28	14	2	2.98

Captain Blood—Sabatini

Grade:	VI	VII	VIII	IX	X	XI	Median grade
Original distribution	1	4	4	6	7	6	9.82
Effect of correction	1	4	4	12	24	42	10.94

Table 6 presents for two books the original distribution of ballots and the distribution obtained from the correction. The original distribution of ballots on *Bow-wow and Mew-mew* is 10, 15, 28, 14, and 2. The median of the reading grades of children reporting on this book according to the original distribution is 4.36. This drops to 2.98 when the corrections are applied to the original distribution. This is a typical result for all books which the *Winnetka Graded Book List* places in the second, third, and fourth grades. For *Captain Blood* the opposite effect results from correction. This is typical of books originally placed in the eighth, ninth, and tenth grades. In the case of *Captain Blood,* the distribution is such as to suggest that if 17,000 ballots were obtained from twelfth-graders, its placement would be even higher.

What is the effect of these corrections on the group of books common to the Winnetka list and Volume I of the *Guide?* Table 7 presents the data for the 159

TABLE 7

EFFECTS OF CORRECTIONS FOR UNEVEN DISTRIBUTION OF BALLOTS ON THE WINNETKA MEDIAN GRADES

Variables

1. Median grades calculated from original Winnetka distributions.
2. Median grades calculated from corrected Winnetka distributions.
3. Grading of Volume I of *Guide* and its rejected list.

	Distribution of books by grade											Mean	*S.D.*
Variables	I	II	III	IV	V	VI	VII	VIII	IX	X	XI		
1	0	0	3	37	72	44	2	1	0	0	0	5.57	.815
2	0	12	21	33	49	28	13	2	1	0	0	5.28	1.518
3	8	16	24	41	43	14	9	2	2	0	0	4.73	1.622

Correlations of grading of *Guide* and rejected list with	
Winnetka original medians	.665
Winnetka medians corrected	.652

books which were found to be common to the two lists.[6] The *Guide* and its rejected list place eight books in the first grade, sixteen in the second, and so on. Their mean grade on the 159 books is 4.73, with a standard deviation of 1.622. The original Winnetka medians yield a mean grade of 5.57, with a standard deviation of .815. The discrepancies are markedly lessened by the corrections. The correction lowers the mean of the Winnetka medians .29 grades and increases the standard deviation until it closely agrees with the *Guide*. On the whole, the correction lowers the Winnetka grading on the 159 books because these books fall for the most part in the lower grades. Although correction for the uneven distribution of ballots results in the marked shifting of the Winnetka grading, it is to be observed from the correlations that the relative order of the books is not disturbed.

[6]In reality, four lists, since grades are available for books which both lists fail to recommend.

IV

EXPLANATION OF DISCREPANCIES: MODE VERSUS MEDIAN

It has just been demonstrated that one of the reasons for the discrepancies between the absolute grading of the *Guide* and the Winnetka list is the failure of the Winnetka list to obtain an adequate sampling of ballots from all the grades of reading ability. Correction was made for this factor. The corrected Winnetka medians are in much closer agreement with the *Guide*. We turn now to a second factor explaining these discrepancies: the statistical treatment of the data.

TABLE 8

DISTRIBUTION OF BALLOTS AFTER APPLICATION OF CORRECTIONS FOR THREE BOOKS SHOWING MODE BY INSPECTION IN THE THIRD GRADE

Book number	Distribution of ballots by grade II	III	IV	V	VI	VII	VIII	IX	Median
1	25	60	39	37	26	5	2		4.18
2	34	47	37	29	15	6	3	4	4.18
3	8	20	14	13	8	2			4.32

Table 8 presents the distribution (after applying corrections for inadequate sampling of ballots) of a sample of three books for which a definite mode by inspection appears in the third grade. For book Number 1 there are 25 ballots from children of second-grade reading ability, 60 ballots from third-graders, and 39, 37, 26, 5, and 2 ballots from fourth-, fifth-, sixth-, seventh-, and eighth-graders. This book is most

widely read and apparently best liked in the third grade. What statistical measure shall be applied to determine the exact best grade of this book? The *Winnetka Graded Book List* uses the median, which places this book in the fourth grade. Similarly, the distribution of ballots on the second and third books indicate that they are most widely read in the third grade, while again the median places them in the fourth. Similar cases are presented in Table 9 for three books showing a mode by inspection in the second grade. Again, the median places them too high.

TABLE 9

DISTRIBUTION OF BALLOTS AFTER APPLICATION OF CORRECTIONS FOR THREE BOOKS SHOWING MODE BY INSPECTION IN THE SECOND GRADE

Book number	Distribution of ballots by grade II	III	IV	V	VI	VII	VIII	IX	Median
1	119	53	42	24	7	1	1	2	3.10
2	25	12	11	6	2				3.25
3	34	27	12	2					3.13

More extreme cases illustrating the opposite tendency of the median are presented in Table 10 for three books showing a mode by inspection in the eleventh grade.

TABLE 10

DISTRIBUTION OF BALLOTS AFTER APPLICATION OF CORRECTIONS FOR THREE BOOKS SHOWING MODE BY INSPECTION IN THE ELEVENTH GRADE

Book number	Distribution of ballots by grade III	IV	V	VI	VII	VIII	IX	X	XI	Median
1	1	2	4	5	10	11	12	17	35	10.26
2			1	2	4	10	18	49	63	10.76
3		2	7	12	22	24	38	35	56	9.82

The medians here place these books in the ninth and tenth grades.

These cases, while not entirely typical, illustrate the errors involved in the median. Books most widely read in the second, third, and fourth grades tend to be placed too high by the median. Books most widely read in the ninth, tenth, and eleventh grades tend to be placed too low by the median. The explanation is simple. Books most suitable for the third grade are often read by fifth- and sixth- and occasionally by eighth- and even ninth-graders. The stringing out of ballots in the upper grades is not compensated by a normal distribution of ballots below the second grade. This explanation is even clearer in the case of the three books of Table 10. These books are in reality adult books. They are often read by eighth-, ninth-, and tenth-graders, and occasionally by still younger children. The conditions under which the Winnetka list was prepared give opportunity for ballots on such books from immature readers and no opportunity for adult readers.

It has been indicated that the distributions are not entirely typical. They are not typical for the reason that clean-cut modes by inspection are not common. In less than one-third of the distributions (after correction) does any grade have the advantage over another grade in number of ballots by as much as 10% of the total ballots. How, then, determine the grade in which a book is most widely read? The authors of the Winnetka list overlooked the possibility of calculating the empirical mode by use of the following formula (3, pp. 60-62):

$$Mode = Mean - \frac{Mean - Median}{c}$$

in which factor c is determined from the formula

$$c = .3309 - \frac{.0846\ (Mean - Median)^2}{8^2 - 9\ (Mean - Median)^2}$$

which for all but extremely skewed distributions equals approximately .33. Mathematically fitting a curve to 159 distributions was out of the question. Accordingly, the above formula, substituting .33 for c, was used to determine the mode.

To illustrate the effect of this second correction the formula may be applied to the data of Tables 8 and 9. For the three books showing a mode by inspection in the third grade and medians in the fourth, the empirical mode yields the following grades: 3.61, 3.59, and

TABLE 11

EFFECT ON WINNETKA GRADING OF CORRECTION FOR THE UNEVEN DISTRIBUTION OF BALLOTS AND FOR THE USE OF THE MEDIAN INSTEAD OF THE MODE

Variables
1. Median grades calculated from original Winnetka distributions.
2. Median grades calculated from corrected Winnetka distributions.
3. Modal grades calculated from corrected Winnetka distributions.
4. Grading of *Guide* and rejected list.

Variables	Distribution of books by grade I	II	III	IV	V	VI	VII	VIII	IX	X	XI	Mean*	S.D.*
1	0	0	3	37	72	44	2	1	0	0	0	5.57	.815
2	0	12	21	33	49	28	13	2	1	0	0	5.28	1.518
3	11	10	26	39	39	26	4	3	0	0	1	4.77	1.606
4	8	16	24	41	43	14	9	2	2	0	0	4.73	1.622

*Winnetka means and standard deviations calculated from a finer grouping.

TABLE 12

CORRELATIONS OF GRADING OF *Guide* AND REJECTED LIST WITH WINNETKA MEDIANS AND MODES FROM ORIGINAL AND CORRECTED DISTRIBUTIONS

Correlations of *Guide* with*	
Original Winnetka medians	.673±.029
Corrected Winnetka medians	.688±.208
Corrected Winnetka modes	.672±.029
Original Winnetka medians with final Winnetka modes	.883±.012

*All correlations calculated from finer groupings than those presented in Tables 11, 13, 14, and 15.

3.97. For the three books showing a mode by inspection in the second grade and medians in the third, the empirical mode yields the following grades: 2.19, 2.60, 2.89.

The effect of this correction on the grading of the 159 books is presented in Tables 11 and 12. For purposes of comparison these tables recapitulate the data of Table 7. The mean of the grades assigned by the *Guide* (variable 4, Table 11) to the 159 books in 4.73, with a standard deviation of 1.622. The mean of the original Winnetka medians (variable 1) is 5.57, with a standard deviation of .815. Successive corrections of the Winnetka grading gradually eliminate these discrepancies until almost perfect agreement is reached in the final correction. That the mode lowers the final mean grade is due to the fact that the 159 books are concentrated for the most part in the lower grades.

It should be pointed out that in a few cases the distributions are so irregular that the mode introduces more error than it corrects for. This is probably true of the one book which is placed in Grade XI according

to the final correction (see variable 3, Table 11). The distribution of ballots after correction on this book is as follows:

Grade:	IV	V	VI	VII	VIII	IX	X
Ballots:	1	2	0	12	6	10	28

The straggling three cases in Grades IV and V and the uneven distribution in Grades VII to X result in a mode of 11.07. There are, of course, no ballots on this book from eleventh graders. This book is placed in the fifth grade by the *Guide*. Although it is tempting to smooth the distribution of ballots, this has not been done.

Apparently more serious errors are involved in the eleven books (see variables 3, Table 11) which the mode places in the first grade. The sum of the distribution of ballots (after correction) on these eleven books is as follows:

Grade:	II	III	IV	V	VI	VII	VIII	IX	X	XI
Ballots:	535	142	132	99	39	15	9	6	4	7

Summation of the ballots in this manner smooths the distribution and frees the mode from the disturbing influences of irregularities in the individual cases. For this distribution the median grade is 2.92 and the modal grade is 1.62. Again, the mode places books in a grade from which no ballots were returned. Whether it is sound to call these first-grade books is an open question. Calling them second-grade books changes the mean grade of the 159 books by only five-hundredths of a grade. We are far enough from the original data.

Further refinements would introduce only slight changes and it is doubtful whether they would bring us closer to the true situation.

Table 12 presents the corresponding correlations. It is to be observed that marked shifting of the means and standard deviations of the Winnetka grading does not alter the correlations. This is important evidence that the corrections applied have not done violence to the data. Tables 13, 14, and 15 display the three most important of these correlations. Tables 13 and 14 are the scatterdiagrams of the *Guide* with the original Winnetka medians and with the final Winnetka modes. According to Table 13 only 37 books are placed in the same grade by the *Guide* and original Winnetka medians. In Table 14, this number is 53. In Table 13 the average difference between the *Guide* and the original Winnetka medians is 23% better than chance arrangement, while in Table 14 the average difference between

TABLE 13

ABBREVIATION OF SCATTERDIAGRAM SHOWING CORRELATION BETWEEN THE GRADING OF THE *Guide* AND THE ORIGINAL WINNETKA MEDIANS

Winnetka Original Medians	*Guide* I	II	III	IV	V	VI	VII	VIII	IX	X	XI
XI											
X											
IX											
VIII					1						
VII						1					
VI			1	6	20	8	6	2	1		
V	1	3	14	27	21	4	2				
IV	6	11	9	8	1	1	1				
III	1	2									
II											
I											

TABLE 14

Abbreviation of Scatterdiagram Showing Correlation between Grading of the *Guide* and Final Winnetka Modes

		Guide										
		I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Winnetka Final Modes	XI					1						
	X											
	IX											
	VIII						1	1	*1*			
	VII					4				1		
	VI			1	3	10	*5*	5	1	1		
	V	1		1	13	*17*	5	2				
	IV	1	3	9	*16*	8	2					
	III		5	*8*	9	2	1	1				
	II	4	*4*	2								
	I	*2*	4	3		2						

the *Guide* and the final Winnetka modes is 44% better than chance arrangement. Table 15 illustrates the tendency of the final mode to lower the grading of books originally placed in the lower grades and to raise the grading of books originally placed in the higher grades.

TABLE 15

Abbreviation of Scatterdiagram Showing Correlation between Grading of Original Winnetka Medians and Final Winnetka Modes

		Final Winnetka Modes										
		I	II	III	IV	V	VI	VII	VIII	IX	X	XI
Original Winnetka Medians	XI											
	X											
	IX											
	VIII											1
	VII							*1*	1			
	VI					13	*26*	3	2			
	V	3		11	32	*26*						
	IV	8	*7*	15	*7*							
	III		3									
	II											
	I											

It will be worth while in addition to the explanations already discussed to point out two additional uncertainties in the grading of the *Winnetka Graded Book List*. In the first place, 25 original ballots are a small number from which to determine the best grade. This is readily demonstrated. The correlation between the grading of the *Guide* and the Winnetka list, using the finally corrected modes, was determined for books having less than 50 ballots originally and for cases having 50 or more ballots. The 75 cases showing less than 50 ballots yield a probable error estimate of .93 of a grade from the Winnetka list. The 84 cases showing 50 or more ballots yield a probable error of estimate of .68 of a grade. That is, the grading of books with less than 50 ballots contains about 36% more error than the grading of books with more than 50 ballots. This is a large difference. It demonstrates that the relative grading of the Winnetka list becomes much more accurate where 50 or more original ballots were obtained.

A second cause of uncertainty in the relative grading of the Winnetka list is that, when children of all grades of reading ability are given the freedom of books of all degrees of difficulty, the resulting distribution of ballots is often spread over the entire range of grades and is occasionally more rectangular than normal. Of the 159 books, the distribution of ballots spreads over the whole 10 grades in 18 cases. In 22 cases the range is 9 grades. In 34 cases the range is 8 grades. The occasional rectangular distribution of ballots (after correction) is well illustrated by the following extreme cases selected from among the 159:

Distribution of Ballots by Grade

	II	III	IV	V	VI	VII	VIII	IX	X	XI
Book No. 1	59	53	54	59	53	25	16	14	3	21
Book No. 2	8	15	14	29	15	21	21	12	14	7
Book No. 3	8	15	6	10	12	17	8	20	7	7
Book No. 4	17	22	10	12	10	8	3	10	3	7

Statistical determination of the best grade of these books by any method whatever is certain to result in serious errors.

The conclusion of these three chapters is that discrepancies in absolute grading are due to inadequate sampling of ballots and to faulty statistical analysis of the Winnetka data. When corrections are applied the disagreements vanish. Instead of being mutually contradictory, grading based on competent opinion and grading based on children's choices are mutually supporting. Chapter VIII will present further evidence on this point.

V

JUDGING THE RELATIVE WORTH OF BOOKS

In this and the following chapters the center of interest turns from the question of agreements concerning the grading of various books to the question of their merit or general worth. Both undertakings list only books of supposedly genuine worth. In the case of Volume I of *A Guide to Literature for Character Training*, 461 titles are recommended according to five degrees of merit, and an additional 399 titles, classified in four levels of demerit, failed of recommendation. The *Winnetka Graded Book List* employs two procedures for selection of the supposedly best books. The first and primary test for inclusion of a book is that at least 25 children must have returned ballots upon it. The number of children reading a book and the number of cities in which it was read determined the popularity index and books are listed according to this index. The second and supplementary test of the worth of a book was the opinion of 13 expert children's librarians. One hundred ten books which three-fourths of these librarians thought unsuitable were excluded from the list. One hundred nineteen which three-fourths of the librarians thought to be of "unquestionable literary merit" were starred. Once more we find the two undertakings attempting to do the same thing by very different methods. The *Guide* relies exclusively on competent opinion to determine the merits and de-

merits of the books subjected to study. The Winnetka list relies primarily upon the choices of children and employs competent opinion only in supplementary and negative fashion. Again, it is asked, do the lists agree? If not, why not? It will be convenient to consider, first, the reliability of the judgments of the librarians and their agreements with the *Guide,* while Chapter VI will consider the question of agreements between judgments of worth by adults and children's choices.

"Just what is 'literary merit' anyhow?" With this question Carleton Washburne and Mabel Vogel sum up the results of their study of the estimates of literary worth by 13 expert children's librarians on the books studied in the preparation of the *Winnetka Graded Book List.*[6] "The reports of these experts varied materially. Out of about 800 titles submitted to them there were only about 100 on which they all agreed. . . . If a group of children's librarians, selected as among the most expert in the United States, differ among themselves as to what books have high literary merit and what ones are trashy, does it not show that none of us are able to set up as yet any final and generally acceptable standard of literary merit?"[7]

The first question at issue is the presence or absence of agreements among the librarians. If it should prove that their estimates are reliable, the second question is whether they agree with the rankings in the *Guide.* To answer these questions there are available two sets of data. Through the courtesy of Mr. Washburne, we

[6]Page 44, original edition.
[7]Pages 43 and 44, original edition.

have the complete record of the original estimates of the librarians on *all* books. Through the courtesy of four librarians, we have the revised estimates on books common to the Winnetka list and those evaluated in preparation of the *Guide*.

The general procedures employed in the preparation of the Winnetka list have been described. It is necessary here to add only a more detailed explanation of how the librarians' judgments were obtained. They were submitted an alphabetical list of all books and were asked to mark each title 1, 2, 3, or 4, indicating whether the book was

(1) *of unquestionable literary merit.*

(2) valuable for the list although not of high literary merit.

(3) not recommended because of low literary merit.

(4) not recommended because of subject matter.[8]

If a book was unfamiliar, they were to indicate the fact with a question mark. These instructions were supplemented by sample titles illustrating the above definitions.

Before presenting the data obtained from these instructions, two comments are pertinent. It will be observed that the librarians were asked to judge two separate things and to judge them on the same scale. Accordingly, the scale of four points is not strictly a scale at all, but calls for four unrelated and sometimes conflicting judgments. If separate estimates of literary merit and of content value were desired, two scales should have been provided. Secondly, the instructions

---

[8]See page 42, original edition

do not insure that the books will be distributed over a reasonable range of steps.

In spite of the unsatisfactory data, statistical treatment yields a very high degree of reliability. The procedure was to consider only those titles upon which at least eight librarians recorded their estimates and to correlate the average of half of these estimates with the other half. This proved to be .862±.012 for a random sample of 200 cases. By Brown's formula the estimates of the thirteen librarians should correlate .925 with a similar group of judges. Further, they should correlate .962 with the judgment of a very large number of similarly competent librarians. Far from being a low reliability, this is most exceptionally high. Indeed, the reliability is so very high as to suggest that we have here, not 13 independent judgments, but 13 judgments derived from a common source.

At the time that an intensive study of the librarians' judgments was undertaken, 125 books upon which at least eight estimates were available were found common to the Winnetka list and those examined in the preparation of Volume I of the *Guide*.

For these 125 books, the chance-half reliability of the original librarian estimates proved to be .825±.091. By Brown's formula the reliability of the 13 estimates is .905. These figures check the reliabilities presented for a random sample of 200 cases.

The average judgment of the librarians correlated against the six levels of merit as given in the *Guide* and the four levels of demerit in a rejected list yields an r of .450±.048. It will be convenient in comparing this

correlation with others to be reported if account is taken of the range of merit involved. For this purpose, we shall use the standard deviation of the ranks according to the *Guide* and its rejected list. This is 1.92 levels of merit. The probable error of estimate is 1.16 levels. That is, a best estimate of the *Guide* from the original librarians' estimates would show discrepancies with the *Guide* which would be less than 1.16 levels in 50% of the cases. We may, however, go beyond the question of obtained agreements with the *Guide* and pose the theoretical question, "How closely would a very large number of librarians agree with a very large number of critics such as those employed in the Institute of Character Research? Correcting the .450 correlation for attenuation, the answer is .520.

The nature of the instructions by which the librarians originally judged the books of the Winnetka list led to a feeling that considerable agreement between the librarians and the *Guide* was being obscured. Accordingly, an attempt was made to obtain revised estimates. A list of the books common to the Winnetka list and Volume I of the *Guide* and their rejected lists was sent to the 13 librarians. They were asked to indicate the best 10% of the books, the next best 15%, the middle 50%, the next poorest 15%, and the poorest 10%. They were instructed to use their own good taste as to the factors determining the best material and to give whatever weight to interest value, literary merit, content, or other values as they individually thought desirable. As a whole, the group proved uncooperative. Only four librarians supplied the revised estimates. But

these are sufficient to give a distinctly higher reliability and a higher correlation with the *Guide*.

As in the study of the original estimates, the first question is that of reliability. Summing two revised estimates and correlating them with the other two yields an r of .692±.034 for 115 cases (10 cases omitted because of incomplete data). While this chance-half reliability is lower than for the original estimates (.825), only four judges are involved instead of thirteen. Correlating the original estimates of two of these judges versus the other two gives a chance-half reliability of only .409±.062 (corrected for errors of grouping). The revised estimates are, accordingly, distinctly more reliable when the number of judgments is considered.

How well do the revised estimates correlate with the *Guide?* That these revised estimates contain new elements is indicated by a correlation of only .69 with the original estimates. Averaging the four revised estimates and correlating them with the *Guide* yields an r of .518±.046 (115 cases). Not only is this correlation slightly higher but, with 10 cases omitted, the standard deviation is smaller, being 1.61 levels. It follows that the probable error estimate is distinctly smaller, being .93 instead of 1.16 levels. Again it is of interest to go beyond the question of obtained rankings and to pose the theoretical question of how well a very large number of librarians would agree with a very large number of critics such as those employed in the Institute. Correcting the .518 correlation for attentuation yields an r of .631.

It would seem to follow from these data that the 'innetka list might well have employed competent inion in a much more primary and positive rôle in-ead of in a supplementary, secondary, and negative le. The data also indicate that, as tested by the com-tent opinion of librarians, the rankings as to merit in e *Guide* are trustworthy. There remains, however, me force in the question, "Just what is literary merit yhow?" Corrected for attenuation, the highest cor-lation is only .63. While librarians agree very well nong themselves and while the readers in the Institute gree among themselves, there is still much disagree-ent between the two groups. Even if the librarians' timates were combined with those of the *Guide,* the mbination would not correlate higher than .68 with e combined opinion of two more similarly competent roups showing similar agreements.

VI

AGREEMENTS BETWEEN MERIT AS ESTIMATED BY COMPETENT OPINION AND BY INTEREST VALUE

In the preceding chapter evidence was presented showing that the judgments of merit by the 13 librarians are highly reliable and that they agree rather well with similar estimates of merit by the readers who prepared the recommendations of the *Guide*. We may now inquire whether the selection of best books by competent opinion gives the same result as selection of best books on the basis of children's choices.

It should be remarked at once that agreements or disagreements here suggest very different conclusions than agreements or disagreements as to grading. If grading by competent opinion and grading on the basis of the reading grade ability of children reading certain books had shown no agreement, the result would have been clearly disastrous to one or the other or both methods of grading. Two methods of grading books cannot give totally different results and both methods be valid. When it comes to selecting superior books, however, the absence of agreements between the literary excellence or general worth of books as judged by mature persons and the appeal of these books to children does not invalidate either basis since they do not purport to measure the same thing. Rather the absence of agreements between judgments of merit and interest value serves to point directly to the necessity of preparing selected lists on the basis of *both* criteria.

In the original studies designed to validate the rankings of books according to the levels of merit given in the *Guide* and its rejected list we made a large number of studies of the interest value data published in the *Winnetka Graded Book List* and its rejected list. No matter how these data were manipulated or what corrections were applied, the same results were obtained. Accordingly, there is reported here only one study. In addition, for the purpose of this volume, we have made a parallel analysis of the relation between merit as judged by the 13 librarians and the various indices of interest value as given in the Winnetka list.

There are no less than five possible indices of children's choices or of interest value recorded in the Winnetka list. Each of these merits a brief description.

1. The index which is given the greatest prominence in the Winnetka list is the number of children reading and reporting on a book. It is, however, not an entirely satisfactory index since it measures availability to an uncertain degree.

2. The number of cities in which a book was read is a possible index of the interest value of a book. Some books were read in 34 cities while others were read in only one city.

3. The Winnetka list reports a "popularity index" which was obtained by multiplying the number of children reading and liking a book by the number of cities in which it was read. The books are listed in order according to this index.

4. In reporting on a book each child checked one of the following statements:

One of the best books I ever read.
A good book, I like it.
Not so very interesting.
I don't like it.

The percentage of children checking the first two statements constitutes a fourth measure.

5. A *fifth measure* of interest was obtained by assigning numerical values of 100, 67, 33, and 0 to the above statements and averaging. The fourth and fifth measures, accordingly, are simply different methods of scoring the same data and should agree very closely.

What do these five indices measure? Table 16 displays their intercorrelations for 156 books which the Winnetka list places in Grade V and for 156 books in Grade VI and on which at least eight librarians' judgments are available. The number of ballots and number of cities are highly correlated to the extent of .817 and .767. Since the popularity index is the product

TABLE 16

INTERCORRELATIONS OF FIVE INDICES OF INTEREST VALUE FOR 156 FIFTH-GRADE BOOKS AND FOR 156 SIXTH-GRADE BOOKS

	Cities	Popularity	Liking	Value
Number of ballots—Grade V	.817	.953	.193	.152
Number of ballots—Grade VI	.767	.956	.106	.140
Number of cities—Grade V		.911	.221	.196
Number of cities—Grade VI		.836	.217	.329
Popularity index—Grade V			.209	.164
Popularity index—Grade VI			.142	.203
Percentage liking—Grade V				.800
Percentage liking—Grade VI				.826
Interest value—Grade V				
Interest value—Grade VI				

of ballots and cities it necessarily correlates highly with its components: .953, .956, .911, and .836. Necessarily, also, the percentage of children liking a book and its average interest value correlate very highly, .800 and .826; but the cross correlations are exceptionally low. Number of ballots, number of cities, and popularity correlate on the average only .189 with the percentage of children *liking* a book and *the* average *interest* value. Obviously, both sets of figures do not measure the same thing.

Which set of figures measures the appeal of the books to the interest of children? Or does neither set measure this factor? A preliminary question is the reliability of the measures, that is, whether they measure anything at all. From the fact that the number of ballots and the number of cities are independent measures and from the fact that they correlate so well, we are safe in inferring that these measures are fairly reliable. Accordingly, the popularity index must also be reliable. The *Winnetka Graded Book List,* however, gives no clue as to the reliability of the percentage-liking and interest-value indices. In the absence of a better measure, the reliability of the percentage liking has been determined indirectly by first calculating the reliability of each percentage and then combining. The estimated reliability turns out to be .72 for grade V and .77 for Grade VI.[9] Presumably it follows that the interest-value index is also fairly reliable.

[9]In detail, the logic is as follows: The reliability of a distribution of scores is defined as the percentage of true variance in the variance of the obtained distribution, or

The most probable interpretation of the very low cross correlations is that the number of ballots returned, the number of cities reporting, and the popularity indices are primarily measures of availability. That availability is certainly involved, at least in part, was demonstrated in our studies of the material of Volume I of the *Guide* by comparing the number of ballots on books in relation to date of publication. Books published prior to 1915 averaged nearly 30 more ballots than books published after 1915. That the number of cities reporting correlates so high with the number of ballots returned also points to availability as the better interpretation. This is not saying that the school and city libraries of the 34 cities which were involved have a limited range of reading material but it can be certain that all the over 9000 books which received at

---

$$r_{11} = \frac{\sigma^2_{\infty}}{\sigma^2}$$

But since

$$\sigma^2_{\infty} = \sigma^2 - \sigma_0^2$$

then

$$r_{11} = 1 - \frac{\sigma_0^2}{\sigma^2}$$

The value σ^2 is the square of the standard deviation of the obtained percentages. The value σ_0^2 may be obtained as follows: The variance of the standard error of any given percentage is

$$\sigma_{0_1}^2 = \frac{p_1 q_1}{n}$$

in which p_1 is the given percentage and $q_1 = 1 - p_1$ then,

$$\sigma_0^2 = \frac{\sigma_{0_1}^2 + \sigma_{0_2}^2 + \sigma_{0_3}^2 + \cdots \sigma_{0_n}^2}{n}$$

least one ballot were not equally available to all the 100,000 children who read various books.

Doubts may also be thrown on the percentage-liking and interest-value indices. The average of the percentage-liking indices is 87.5. That is, only about 13% of the children checked either the statement "Not so very interesting" or "I don't like it" in reporting on a book. Further, the introduction to the *Winnetka Graded Book List*[10] reproduces a curve showing that the percentage of children liking a book did not change from grade to grade as might be expected. In spite of these two uncertainties, the author is inclined to the judgment that the percentage-liking and interest-value indices are more valid measures of the true appeal of books than either the number of ballots, the number of cities, or the so-called popularity indices.

With this provisional interpretation in mind, we consider the correlations between the various measures of interest value and competent opinion. These are reported in Table 17. The first two columns give the correlations between the indices and the estimates of the 13 librarians. Columns three, four, and five record the correlations between the indices and the levels of merit according to the *Guide* and its rejected list. The 24, 53, and 34 books are all those placed in Grades IV, V, and VI by the Winnetka list among the 125 books common to the *Guide* and the Winnetka lists. On the whole, competent opinion correlates positively with number of ballots, cities, and popularity, the values ranging from —.006 to .540, the weighted averages

[10]*Pages 20 and 21, original edition.*

being .246, .167, and .213. On the whole, competent opinion correlates zero or negatively with the percentage-liking and interest-value indices, the r's ranging from —.370 to .310, the weighted averages being —.254 and —.207. On the whole, the correlations between the indices and merit as recorded in Volume I of the *Guide* and its rejected list are slightly more positive than the correlations between the indices and librarian estimates of merit. This is due to the fact that the materials of Volume I are fairy tale, myth, and legend. Poor books in the eyes of adults which make strong appeals to children are much less common in the field of the fairy tale than they are in the whole area of children's books.

These data tend to confirm the interpretation of the intercorrelations of the Winnetka indices. That competent opinion gives slightly positive correlations with number of ballots, number of cities, and popularity may be due to the dependency of availability upon the recommendations of librarians and others who prepare lists of books for purchase by school and city libraries.

TABLE 17

CORRELATIONS OF COMPETENT OPINION WITH VARIOUS INDICES OF INTEREST VALUE

	Librarians		The Guide			
Indices	156 books Grade V	156 books Grade VI	24 books Grade IV	53 books Grade V	32 books Grade VI	Weighted Average Correlation
Number of ballots	.282	.159	.142	.377	.351	.246±.031
Number of cities	.158	.041	—.006	.421	.540	.167±.032
Popularity index	.260	.080	.035	.425	.417	.213±.031
Percentage liking	—.295	—.384	.026	—.076	.077	—.254±.031
Interest value	—.370	—.377	—.069	—.103	.310	—.270±.030

The negative correlations between competent opinion and percentage-liking and interest-value were a complete surprise. There is no way of interpreting these data so as to yield the conclusion that there is any substantial correlation between what children like and what adults approve. This does not prove, but it does warrant the inference, that lists of best books should be selected on the basis of both competent opinion and interest value.

VII

EXPERIMENTAL RESULTS: THE CORRELATION BETWEEN JUDGMENTS OF MERIT AND INTEREST VALUE

In this and the following chapter additional data are presented which confirm and supplement the conclusions of the previous chapters.

Early in our studies of the fairy-tale material in the preparation of Volume I of the *Guide,* an exploratory experiment was set up for the purpose of testing the reliability of three methods of measuring the interest value of individual stories. Individual stories were selected for study since these could be mimeographed and read by large numbers of children without excessive expense. It was planned after testing the reliability of the three measurements of interest value to undertake further studies of (1) the correlation between judgments of merit and interest value and of (2) the correlation between grade placement and interest value. The further precise studies of the correlation between judgments of merit and interest value, most unfortunately, were never undertaken. The exploratory experiment, however, provides data which are worthy of presentation in this connection. The present chapter summarizes these data, while Chapter IX presents the results of a more precisely controlled study of the correlation between grade placement and interest value.

The exploratory study was designed to test the reliability of three methods of measuring the interest

value of individual stories. For the purpose of the experiment, 36 stories were selected and mimeographed. These stories were selected from among 157 which were the first ones to be subjected to systematic analysis by the staff of the Institute of Character Research early in 1926. The selection was primarily for moderate length and variety of types. As then judged by the staff all of these stories were of average or very superior merit, but later evaluations in the light of all the fairy-tale material showed that the entire range of merit was included. The stories were divided into two groups of 20 stories, each group having four stories common to the other. Each of these groups was further divided into four sets of five stories each. These divisions were purely mechanical, the effort being to have each set contain the same amount of reading material. Within each set, the stories were arranged in three different orders to avoid any position error. One group of 20 stories was read by 115 children from the third to the sixth grade inclusive in the schools of Spencer, Iowa, under the supervision of Superintendent J. R. McAnelly. The other group was read by 208 children from the third to sixth grade inclusive in the schools of Mason City, Iowa, under the supervision of Superintendent F. J. Vasey.

The children made three reports on each story as to its interest value. Method I was as follows: After each story was read, they were asked to indicate its interest value on a scale of five statements thus:

. . . . One of the best stories I ever read.
. . . . A good story, I like it.
. . . . I neither like nor dislike this story.
. . . . Not so very interesting.
. . . . I dislike this story very much

Method II consisted of indicating the best-liked and the least-liked story out of each set of five. When the reading of the group of 20 stories was completed the children were given a list of the stories and were asked to indicate the three best-liked and three least-liked stories of the 20. These reports constituted Method III. A uniform procedure was used throughout, the teachers being particularly warned not to indicate in any way their own attitude toward any story. The children were given all the time they desired.

The reports from each of the three methods were tabulated separately. Responses to Method I were assigned values of 5, 4, 3, 2, and 1 for each of the five steps from liking to disliking. On Methods II and III, a value of three was given to each story rated as "best-liked," a value of one was given to each story rated as "least-liked," and a value of two to all other stories. These scores were summed and divided by the number of children reporting in order to obtain an average. The Mason City and Spencer data on the four stories read in both schools were combined and the 36 stories treated as a unit.

The reliability of each method for each grade was determined by dividing the children into chance-half groups and applying Brown's formula. These are presented in Table 18. From the table, the three methods are seen to be of nearly equal reliability, except in the third grade where the checking of the three best-liked and three least-liked stories out of 20 has the advantage. The very high reliabilities, all over .95, obtained by combining the data for the fourth, fifth and sixth grades

are to be noticed. The degree to which the three methods measure the same thing was determined by calculating the intercorrelations for the combined data of the fourth, fifth, and sixth grades. Method I with II gives a correlation of .84, I with III gives .88, and II with III gives .94. A combination of these three measures should correlate .96 with three more methods yielding similar intercorrelations.

The procedure itself and the internal consistency of the data are the best evidences of the validity of these measures of interest value. The theoretical validity (square root of the reliability) is .98. As a supplementary test eight literary critics in the Institute of Character Research, who had had six months' intensive training in the judging of these materials for their general worth, were asked to estimate the interest values of these stories. As an aid they were given an adaptation of Uhl's (7) standards of interest value which placed the emphasis on "dramatic, exciting, and adventurous action" and on "clearly drawn characters portraying kindness and faithfulness." The reliability of their judgments proved to be .64. These estimates of interest value correlated .590 with a pool of the three methods of measurement. Corrected for attenuation this becomes .752.

Before presenting the correlation between the pooling of these measures of interest value and adult estimations of their merit, it should be remarked once more that the data were collected for the purpose of testing the reliability of the three methods of measuring interest values and not for the purpose of testing the

TABLE 18

RELIABILITY OF CHILDREN'S STATEMENTS AS TO THE INTEREST VALUE OF STORIES FOR EACH GRADE AND EACH METHOD OF MEASUREMENT

N=36.

Grade	Method I	Method II	Method III
III	.656	.594	.851
IV	.877	.883	.871
V	.900	.949	.864
VI	.924	.931	.915
IV-V-VI	.962	.974	.952

relationship between merit and interest values. It follows that the data fall short of what might be desired in several respects. First, the number of stories involved is too small. Secondly, while the sample of stories contains a wide variety of tales, it is slightly overweighted with fables. Thirdly, the stories are spread over four school grades. There are eight third-grade stories, twenty-one fourth-, five fifth-, and two sixth-grade stories as judged by the staff readers. A more desirable selection would have been confined to stories all of the same grade. Fourthly, in the selection of stories no effort was made to insure a representative sampling of the various levels of merit with the result that the range of merit involved is too large.

Two measures of merit are available. One consists of the degrees of merit assigned by the *Guide* supplemented by the judgments on rejected stories of two readers who had charge of making all the final adjustments in the rankings of stories as published in the *Guide*. The second consists of special estimates made by the entire staff of readers. No direct measure of the

reliability of the first measure is available, but it is probably below that of the second, which is .85.

Any obtained correlation between interest values and estimated merit will depend largely on the presence or absence of differences in interest value and merit. It may be inferred from the very high reliabilities that as to interest value these stories represent genuine differences. The distribution according to the first measure of merit is more than adequate with a standard deviation of 2.09 in comparison with an estimated standard deviation of 1.60 for all the 6000 stories which were evaluated. With the presence of genuine differences in interest value and merit the stage is set to yield maximum correlation. The pooling of all three measures of interest value and of responses from Grades IV to VI correlates $.305 \pm .102$ with the special estimates of the entire staff and $.254 \pm .106$ with the rankings as published in the *Guide* supplemented by the judgments of two readers. The most significant of these correlations falls short of being three times as large as its probable error. In view of the large differences involved and the unreliability of the correlations, the data indicate that for these 36 stories there is no relationship between merit as judged by competent adults and the interest value of the stories for children. It is, of course, extremely hazardous to generalize from only 36 stories as to the true relationship for the some 6000 tales in the field which were carefully examined by the staff of the Institute.

VIII

EXPERIMENTAL RESULTS: GRADE PLACEMENT AND INTEREST VALUE

The present chapter reports experimental data employing a limited number of individual stories to supplement the material of Chapters II to IV concerning grade placement. Fundamentally, the issue involved is whether competent persons can place reading materials at their correct absolute level in the several school grades.

The procedure employed took its cue from certain results of the exploratory study of methods of measuring interest values which hinted that stories best-liked in one grade were not the ones best-liked two or three grades removed. Twenty tales were selected for study. All were of approximately the same length, the same level of merit, and all were distinctly fairy tales. They consisted of four stories presumably most suitable for second-grade children, four stories for the third grade, four for the fourth, four fifth- and four sixth-grade stories. This grading was carefully verified by the staff of readers of the Institute of Character Research at the University of Iowa and any story which fell half way between two grades was excluded from the selection. From other studies the probable error in the relative grading of these stories is estimated at four tenths of a grade. That is, relative to every other story, the chances are that not more than two or three of the twenty should be moved up or down one grade to con-

form to a theoretically true relative grading. Accordingly, for the purpose of this chapter, we shall assume that the relative grading is highly accurate. It is not assumed, however, that the absolute grading is accurate since the problem here is to determine whether the group of twenty tales as a whole should be moved up or down in its grade placement to conform to a true absolute grading.[11] Accordingly, we shall designate the tales by the non-committal letters *A, B, C, D,* and *E* and shall proceed to test whether they belong in Grades I to V, or II to VI, or III to VII.

Since the preliminary study showed little difference in the three methods of measuring interest values, the simplest method was used. This consisted of asking children to select the one best-liked and one least-liked story out of a set of five. Each set of five contained an *A, B, C, D,* and *E* story. Eight sets were prepared in such a way that each story was directly compared with seven others. Within each of these sets the stories were arranged in all possible orders to avoid any position error. Uniform instructions, as in the previous experiment, were sent to all teachers immediately in charge of the reading. The stories were read during the middle of the school year by 788 children from the second to the sixth grades, inclusive, in the schools of Mason City, Iowa, again through the courtesy of Superintendent F. J. Vasey.

The resulting data were subjected to the same analysis as in the first study. The reliability of the interest

[11]See Chapter II for the distinction between relative and absolute grading.

indices based on the judgments of second-grade children is only .29, but for Grades III, IV, V, and VI, the reliabilities are .86, .88, .96, and .98, respectively. The indices obtained from the judgments of second-grade children correlate .504 with those of third-grade children and then drop to .012, .018, and .027 when correlated with the data supplied by fourth-, fifth-, and sixth-grade children. Similarly, the indices obtained from third-grade children correlate .781 with those of fourth-grade children and then fall to .612 and .564 when compared with fifth- and sixth-grade indices. The fourth-grade indices correlate .920 and .802 with fifth-and sixth-grade indices, and the fifth-grade indices correlate .955 with those of the sixth grade.

In order to trace the trends in the interest values of individual stories the indices obtained from each school grade of children were translated into sigma deviations with the mean set at zero and the standard deviation at one. The resulting data are displayed in Table 19. The Roman numerals across the top of the table indicate children in Grades II to VI. The capital letters at the left indicate the supposedly most suitable grade for each story, *A* being second grade, *B* third grade, etc. The subscripts *a, b, c,* and *d* distinguish the four stories of each kind. Story *Aa* stands 1.2 sigma deviations above the mean of all stories according to second-grade children. Only one story is more interesting to second-graders. The standing of story *Aa* among third-grade children, however, is 1.6 sigma deviations below the mean of all stories, and only one story is more disliked. Here, in only the next adjacent grade, is a fall from

TABLE 19

INTEREST INDICES IN TERMS OF SIGMA DEVIATIONS FOR TWENTY STORIES BASED ON EXPRESSIONS OF INTEREST BY CHILDREN IN GRADES II TO VI

Stories	Interest indices in Grades II to VI				
	II	III	IV	V	VI
Aa	1.2	—1.6	—2.6	—2.5	—2.0
Ab	0.0	—0.5	—1.2	—1.8	—2.0
Ac	0.0	—0.8	—0.8	—1.4	—1.7
Ad	—1.0	—1.8	—0.5	—1.3	—1.5
Average	0.1	—1.2	—1.3	—1.8	—1.8
Ba	0.0	1.7	0.7	0.3	0.2
Bb	0.0	0.7	0.3	0.3	0.4
Bc	0.0	0.0	—0.5	—0.3	—0.2
Bd	1.1	0.3	—0.6	0.2	0.1
Average	0.3	0.7	0.0	0.1	0.1
Ca	0.0	0.6	0.8	0.0	0.0
Cb	0.6	1.1	1.3	0.8	0.6
Cc	2.7	1.3	0.7	1.0	0.7
Cd	—0.4	0.3	0.3	0.2	0.4
Average	0.7	0.8	0.8	0.5	0.4
Da	—0.4	—0.2	1.1	1.3	1.5
Db	—1.3	—1.1	0.3	0.1	0.5
Dc	1.2	0.7	0.9	1.1	1.3
Dd	—0.8	—1.1	0.6	0.8	0.4
Average	—0.3	—0.4	0.7	0.8	0.9
Ea	—1.7	0.0	0.1	—0.3	0.3
Eb	—0.1	—0.3	—1.2	0.2	—0.2
Ec	0.2	1.7	1.4	1.1	0.9
Ed	—1.3	—0.8	—1.1	0.1	0.1
Average	—0.8	0.1	—0.2	0.3	0.3

grace with a vengeance. Similarly, all of the *A* or supposedly second-grade stories stand higher in the eyes of second-grade children than in the eyes of third-grade children. The averages tell the tale of the fall in the market value of these stories starting with 0.1 sigma in the second grade and dropping to —1.2, —1.3, —1.8, and —1.8 sigma deviations in the sixth grade. The fig-

ures for the other stories are interpreted in like manner. On the average, the four *B* or supposedly third-grade stories reach their peak of interest in the third grade. On the whole, the four *C* stories stand highest in the eyes of third- and fourth-grade children. They are almost as well liked by second-grade children and distinctly less well liked by fifth- and sixth-graders. Both the groups of *D* and *E* stories show a steady and marked rise in interest value from the earlier to the later grades.

Save only the indices obtained from second-grade children, the standard errors of the indices are comparatively small, ranging from .37 to .14 sigma. But since the indices in Grades III to VI are highly correlated, the standard errors of the differences[12] are even smaller, ranging from .32 to .08. Taking each individual story at a time, the significance of all the differences in its standing in the five grades was tested. Forty-five differences proved to be three or more times as large as their standard errors. Fourteen of the 20 stories show statistically significant changes in their interest value depending on the grade in which they are read. We may also calculate the significance of differences between differences. For example, the difference between story *Aa* and *Ba* in Grade II is 1.2 sigma, while in Grade III it is —3.3 sigma. The difference between these differences is 4.5 points. That is, if both are started from the same point the total divergence becomes 4.5 points. This difference and many more of the same type are well beyond three times as large as their standard errors. By this method of analysis, the

[12] S.E. difference $= \sqrt{\sigma^2_1 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2}$

trends of each of the 20 stories assume statistical significance.

We turn now to the primary problem, that of absolute grading. Do these stories which the readers in the Institute of Character Research place in Grades II, III, IV, V, and VI belong in these grades? Or should they be moved up or down a grade or more? To test this question, it will be convenient to start with three alternative hypotheses and an assumption. The assumption is that the school grade showing the highest interest value should be the best grade for a story and that the indices should show a progressive falling off as one moves away from the grade of highest interest value. Our hypotheses will be designated *X, Y,* and *Z*. Hypothesis *X* is that the *A, B, C, D,* and *E* stories should be placed in Grades I to V. Hypothesis *Y* is that these stories belong in Grades II to VI. Hypothesis *Z* asserts that these are third-, fourth-, fifth-, sixth-, and seventh-grade stories. Which of these hypotheses best fits the data?

Consider first the *A* stories. According to hypothesis *X* these should be first-grade stories and should show declining interest indices from the first to the later grades. According to hypothesis *Y,* these are second-grade stories and should show falling indices beyond Grade II. According to hypothesis *Z,* these should be third-grade stories and should show a rise in interest value to Grade III and then a decline. Hypotheses *Y* and *Z* state precisely opposite trends for the *A* stories between grades II and III. Turning to the data, we find that these four stories behave according

to hypothesis Y. Continuing this testing process with the B, C, D, and E stories, we have the following results. Out of 16 tests as between X and Y, the data favor hypothesis Y in 8 instances and X in 5, while in 3 instances the test is uncertain due to the absence of a trend. As between Y and Z, out of 16 tests, the data favor Y in 12 and Z in 4 instances.

Consider, secondly, the A and B stories. According to hypothesis Y the A stories should fall and the B stories rise in interest value between Grades II and III, while according to X both groups should show declining indices. Here there are 16 tests. Four are uncertain. Of the 12 decisive tests the data fit hypothesis Y in 8 and X in 4 instances. Continuing this type of comparison as between X and Y there are 96 such tests in the table. Of these, 32 are uncertain, 45 favor Y and 20 favor X. Similarly, as between Y and Z, out of 96 tests, 40 are uncertain, 42 favor Y and 14 favor Z. Combining these data with the 32 tests of single stories, we have the following percentages. Of decisive tests as between X and Y, 32% favor X and 68% favor Y. As between Y and Z, 75% favor Y and 25% favor Z. These percentages differ from a chance 50-50 division by more than five times their standard errors. Hypothesis Y, that the A, B, C, D, and E stories are second-, third-, fourth-, fifth-, and sixth-grade stories, fits the data more than twice as well as hypothesis X and three times as well as Z. That X and Z so nearly balance each other tends also to confirm the advantage of hypothesis Y. A further check is to test whether hypothesis X or Z best fits the data. Here there are 192 possible

tests, of which 83 are uncertain. Of the 109 decisive tests, 54% favor X and 46% favor Z. That these two hypotheses are equally satisfactory points to the superiority of the intermediate hypothesis Y. It should be noted also that as between X and Z the uncertain tests constitute 44% of all tests, while as between X and Y and between Y and Z the proportion of uncertain tests is only 34%.

Again, it is hazardous to generalize from so small a number of cases. The data, however, are important confirmation of the conclusion that competent opinion can be relied upon to place childrens' reading materials correctly not only on a relative but also on an absolute grade scale.

IX

SUMMARY, INTERPRETATIONS, AND PROBLEMS

As stated in Chapter I, the function of studies of this monograph is to mediate between two opposed procedures employed in the preparation of graded lists of superior reading materials for children. One procedure, adopted by the *Winnetka Graded Book List,* relies almost entirely on data collected directly from children. The other, adopted by the volumes of a *Guide to Literature for Character Training,* relies exclusively on competent opinion. Agreements and disagreements between the results of these two approaches have been tested as to (1) grade placement and as to (2) general worth.

The results of the studies of the placement of books in the several school grades are rather decisive and unambiguous and may be stated somewhat categorically. The original grades reported in the Winnetka list agree very well with those reported in the volumes of the *Guide* and with competent opinion in general as to relative grading but do not agree as to absolute grading. That is, the correlations between the Winnetka grading and competent opinion are high, while the average grade and range of grades assigned are in disagreement. Books which competent opinion places in Grades I, II, and III tend to be located in Grades IV and V by the Winnetka list, and books which competent opinion places in Grades VIII, IX, and X tend to

be located in Grades VII and VIII by the Winnetka list. These discrepancies are due to the failure of the Winnetka list to obtain an adequate sampling of ballots from the earlier and later grades and to the failure of the Winnetka list to apply the preferable measure of central tendency. Corrections were applied for both of these factors to 159 books common to the Winnetka list and Volume I of the *Guide*. When the corrections are applied, the discrepancies in absolute grading disappear. Within the ever-present limitations imposed by ncessary errors of measurement, reliance upon competent opinion or reliance upon data obtained directly from children should give the same grading. While the actual grading reported by the Winnetka list is faulty, the essential ideas involved are entirely sound and worthy of special commendation. The conclusions are supported by an experimental study of 20 individual stories.

The second concern of these studies was to test agreements between two methods of selecting superior books for children. The *Winnekta Graded Book List* purports to make its selection primarily on the basis of children's interests supplemented by the competent opinion of 13 librarians; the volumes of the *Guide* rely exclusively upon competent opinion. Since the introduction to the Winnetka list leaves the impression that the librarians' judgments are unreliable, a necessary preliminary point which had to be considered was the agreement between the two sets of competent opinions. Analysis of the data, however, shows that the reliability of the librarians' judgments is most exceptionally high.

Further, the estimates of merit by the librarians show substantial agreements with the rankings of merit according to the *Guide*.

The evidence is that the estimates of merit by competent opinion do not agree with the appeal of the books to children. The data show very low or negative correlations between competent opinion and the percentage of children liking a book and average interest values and only slightly positive correlations between competent opinion and number of ballots returned, cities reporting, and the so-called popularity index. The positive correlations are interpreted as due to the probability that these three indices measure availability more than they measure the appeal of the books. An experimental study of 36 individual stories also indicates that there is very little correlation between what children like and what adults approve.

These results are of crucial importance in the selection of superior reading materials. Librarians, teachers, and parents rightly insist that the reading material of children be free from objectionable content. Presumably, children will not read what is uninteresting. Close agreements between competent opinion and children's choices would have meant that the preparation of selected lists might employ either method or both combined. The absence of agreements does not prove, but it does create the inference, that *both* competent opinion *and* the choices of children must be taken into consideration.

This conclusion is critical both of the volumes of *A Guide to Literature for Character Training* and of

the *Winnetka Graded Book List.* While the judgments of the librarians indicate that the readers in the Institute of Character Research have done a reasonably good job in estimating literary merit and content values, the absence of agreements with children's choices indicates that this is only half the battle. While the general plan of the Winnetka investigation is sound, the translation of the plan into concrete terms leaves much to be desired. Instead of presenting the popularity index as the best single measure of the general worth of a book, an index based upon the librarians' estimates and upon the percentage liking or interest values should have been employed. The number of ballots returned on a book would then have served in negative rôle to eliminate books whose appeal to children was inadequately measured.

No one should draw the conclusion that these studies point the way to methods of selecting superior reading materials. There is a need, first, for a good many studies of the interest value of books and stories under controlled conditions. There is a need, secondly, for further testing and refinement of the methods of obtaining expressions of interest from children on a large scale. The percentage-liking and interest-value indices developed by the Winnetka study are a start in this direction. The author is inclined to believe that these are more valid measures of the appeal of books than either the number of ballots returned or the number of cities reporting on the popularity indices. But, just *how* valid the percentage-liking and interest-value indices really are is a question which these studies do not even

touch. Thirdly, there is a need for further refinements in methods of judging the worth of reading materials by competent persons. The Institute of Character Research has made a notable contribution in this direction. It seems to the author, however, that the Institute has sacrificed validity for reliability. A greater variety of points of view and background while giving less reliable estimates would improve validity. The same criticism may be applied even more aptly to the estimates of the 13 librarians. These are so very highly reliable as to suggest either the absence of healthy differences of opinion or that these judgments are not independent at all but derived from some common source. Finally, there must be many studies of the function of supplementary reading materials in relation to the specific and more general aspects of the whole educative process. Is it true, as suggested in the *Winnetka Graded Book List,* that a wide variety of interesting, worth while, and properly graded books will aid materially in solving many of the problems in teaching children how to read? Is it true, as asserted in Volume II of *A Guide to Literature for Character Training,* that the moral values presented in the story will carry over into life?

REFERENCES

1. Carnegie Library of Pittsburg. Books for grades one to eight. Pittsburgh, Pa.: Carnegie Library, 1907.
2. Elementary School Library Committee of the National Education Association. Graded lists of books for children. Chicago: American Library Asso., 1922.
3. KELLEY, T. L. Statistical method. New York: Macmillan, 1923. Pp. xii+390.
4. STARBUCK, E. D. & SHUTTLEWORTH, F. K. A guide to literature for character training: Vol. 1. Fairy tale, myth, and legend. New York: Macmillan, 1928. Pp. xii+389.
5. STARBUCK, E. D., *et al.* A guide to books for character: Vol. II. Fiction. New York: Macmillan, 1930. Pp. x+579.
6. State of Wisconsin. List of books for school libraries in the State of Wisconsin. Madison, Wis., 1926.
7. UHL, W. L. Scientific determination of the content of the elementary school course in reading. University of Wisconsin, 1921.
8. WASHBURNE, C., & VOGEL, M. The Winnetka graded book list. Chicago: American Library Asso., 1926. Pp. 286. (Republished under the title *What children like to read.* Chicago: Rand McNally.)

ÉTUDE CRITIQUE DES MEILLEURS LIVRES POUR LES ENFANTS

(Résumé)

On a fait cette étude dans le but d'évaluer les résultats de deux méthodes différentes employées dans la préparation de listes progressives de morceaux supérieurs de lecture pour les enfants. Une méthode, adoptée par la *Winnetka Graded Book List,* dépend presque entièrement de données fournies directement par les enfants. L'autre, adoptée par les volumes d'un *Guide to Literature for Character Training,* dépend exclusivement de l'opinion des adultes compétents. On se demande si des méthodes si différentes donnent les mêmes résultats.

La liste Winnetka et le Guide recommandent une certaine année scolaire comme la plus adaptée à chaque livre. Bien que la corrélation entre les deux recommandations soit élevée, on a trouvé de grandes différences dans la position des années scolaires moyennes et la variation des années employées. On a trouvé que ces différences ont été dues à deux fautes dans la compilation des données Winnetka. Quand on les a corrigées, les différences ont disparues. Employée correctement, la méthode de dépendre des données fournies par les enfants ou la méthode de dépendre de l'opinion compétente donnent le même résultat. Cette conclusion soutient la validité des deux méthodes.

Le deuxième but de cette étude a été de tester les différences dans le choix des livres supérieurs. La liste Winnetka les choisit premièrement sur la base des intérêts des enfants; les livres du Guide dépendent exclusivement de l'opinion des adultes compétents. L'évidence montre que les jugements du mérite par l'opinion compétente ne s'accordent pas avec l'attrait des livres pour les enfants. Un bon accord de l'opinion compétente et des choix des enfants aurait signifié que la préparation de listes choisies de morceaux de lecture pourrait employer l'une ou l'autre méthode. L'absence des accords ne prouve pas mais fait supposer que l'on doit considérer et l'opinion compétente et les choix des enfants. Cette conclusion critique les volumes d'un *Guide to Literature for Character Training* et ceux de la *Winnetka Graded Book List*.

SHUTTLEWORTH

EINE KRITISCHE UNTERSUCHUNG ZWEIER LISTEN DER BESTEN BÜCHER FÜR KINDER

(Referat)

Das Ziel dieser Untersuchung war, die Resultate von zwei gegensätzlichen Verfahren, die in der Zubereitung von rangmässig geordneten Listen höherstehenden (superior) Lesematerials für Kinder benutzt worden sind, abzuschätzen. Das eine Verfahren, welches bei der Verfertigung des Winnetka Graded Book List gebraucht worden ist, verlässt sich fast vollkommen auf Daten, die direkt an Kindern gesammelt worden sind. Das andere, das in den Bänden des "Guide to Literature for Character Training" (Wegweiser zur Literatur zur Bildung des Karakters) Gebrauch gefunden hat, stützt sich ausschliesslich auf die Meinung sachkundiger Erwachsener. Es wird die Frage gestellt, ob solche verschiedene Methoden die selben Resultaten liefern.

Sowohl in der Winnetka Liste wie in dem Wegweiser wird im Zusammenhang mit jedem Buche eine besondere Schulklasse als diesem Buche be-

sonders angeeignet anempfehlen. Obwohl die Korrelation zwischen den zwei Gruppen von Empfehlungen eine hohe ist, zeigten sich doch grosse Unterschiede in der mittleren Stellung und in dem benutzten Umfang (range) der Schulklassen. Diese Widersprüche konnten auf zwei Fehlverfahren bei der Verfertigung des Winnetka Materials zurückgeführt werden. Nachdem Korrigierungen angewendet worden waren, verschwanden die Widersprüche. Wird sie richtig angewendet, so leistet die Methode des Vertrauens auf an Kindern gesammeltes Material das selbe Resultat, wie die des Vertrauens auf die Meinung sachkundiger Erwachsener. Dieser Befund unterstützt die Gültigkeit beider Verfahren.

Die zweite Beschäftigung dieser Untersuchung bestand darin, dass man die Uebereinstimmung bei der Auswahl höherstehender Bücher prüfte. In der Winnetka Liste wird die Auswahl vorherrschend durch die Interessen von Kindern bedingt. In den Bänden des Wegweisers wird das Vertrauen, wie gesagt, vollkommen auf die Meinung sachkundiger Erwachsener gestellt. Es zeigt sich, dass die Werturteile nach sachkundiger Meinung mit der Anziehungskraft, die die Bücher auf Kinder ausüben, nicht übereinstimmen. Eine innige Uebereinstimmung der sachkundigen Meinung mit den Auswahlen von Kindern hätte angedeutet, dass zur Verfertigung von Listen von ausgewähltem Lesematerial für Kinder sowohl das eine wie das andere Verfahren geeignet ist. Der Mangel an Uebereinstimmungen beweist nicht, weist aber doch darauf hin, dass sowohl sachundige Meinungen wie die Auswahlen von Kindern in Betracht gezogen werden müssen. Dieser Schluss stellt sich sowohl den Bänden des "Guide to Literature for Character Training" wie dem "Winnetka Graded Book List" kritisch gegenüber.

SHUTTLEWORTH

$7.00 per volume · This Issue $3.00 · MONTHLY · Two volumes per year · May-June, 1932 · Volume XI, Nos. 5 & 6

GENETIC PSYCHOLOGY MONOGRAPHS

Child Behavior, Animal Behavior, and Comparative Psychology

MEASURING HUMAN ENERGY COST IN INDUSTRY:

A General Guide to the Literature*

From the University of Chicago

By

RICHARD M. PAGE

*Recommended for publication by W. V. Bingham and Arthur W. Kornhauser, accepted by Carl Murchison of the Editorial Board, and received in the Editorial Office April 30, 1931.

Worcester, Massachusetts

Entered as second-class matter December 1, 1925, at the postoffice at Worcester, Mass., under Act of March 3, 1879

PREFACE

This work represents a careful and thorough review of the literature on respiratory exchange determination directed especially towards enlarging our information on the effects of muscular activity, recovery from such activity, and fatigue factors on the oxygen consumption of individuals. It differs from other available reviews in that the author has made no attempt to enlarge or complicate the work by introducing pathological conditions, or other phases of metabolic studies not directly pertaining to his principal theme. The section which deals with methods is especially worthy of note in that the author has spared no attempt to include all of the references pertaining to methods involved in this type of technique. The physiology introduced is of an elementary nature since the author has compiled this work for psychologists and industrial workers who are not completely trained in physiological methods.

The work gives promise of being a valuable reference guide to industrial workers and psychologists interested in quantitative measurements of activity and efficiency of the human body as determined by the O_2 consumption and CO_2 excretion methods.

M. M. Kunde

Department of Medicine
The University of Chicago
Chicago, Illinois

CONTENTS

PREFACE 323

INTRODUCTION 327

PHYSIOLOGICAL FOUNDATIONS 331
- Theory of Respiratory Exchange Determination . 331
- Fatigue and Energy Cost 332
- Literature on Fatigue Measurement 333
- General Physiology 334
- Basal Metabolism 337
- General Metabolic Exchange and Energy Metabolism 338
- Respiratory Physiology and Blood Chemistry in Relation to Muscular Work 341
- The Biochemistry and Dynamics of Muscle Action 349
- The Physiology of Muscular Work 353
- The Respiratory Quotient 359
- Calorie Computation 363
- Factors Affecting Metabolism 367
- Physiological Factors Subject to Experimental Control 378
- Mental Work as a Physiological Factor . . . 388
- Environmental Factors Affecting Metabolism . 392
- *General References on Factors Affecting Metabolism* 398

APPARATUS AND METHODS 400
- General Manuals on Apparatus and Methods . 400
- Incidental and Non-Quantitative Indexes of Metabolic Rate 402
- *Direct Calorimetry* 406
- Indirect Calorimetry: Open Circuit Methods, Gasometer Type 407
- Open Circuit Methods, Douglas Bag Type . . . 411
- The Waller Simplification of Douglas Bag Method 415
- Criticism and Defense of Waller Method . . 420
- Gas Analysis Under the Waller Method . . 435
- Other Simplified Methods of Gas Analysis . . 438
- Other Open Circuit Methods of Metabolic Rate Determination 440
- Total Pulmonary Ventilation as an Index of Engery Consumption in Light Work 444

Closed Circuit Methods for Determining Oxygen Consumption 447
Closed Circuit Methods for Determination of Oxygen Consumption and Carbon Dioxide Output 453
Graphical Methods in Respiratory Exchange Determination 457
Breathing Apparatus for Respiratory Exchange Experiments 459
Accessory Apparatus in Respiratory Exchange Experiments 461

III. APPLICATIONS AND RESULTS 464
Uses of Respiratory Exchange Determination in Industry 464
Laboratory Studies of the Components of Industrial Work 464
Energy Cost of Industrial Occupations 470
Energy Cost of Walking 473
The Study of Industrial Working Conditions . 476
The Effect of Noisy Working Conditions . . 479
Rest Pauses 481
Metabolic Tests of Physical Efficiency . . . 483

BIBLIOGRAPHIC NOTES 485

BIBLIOGRAPHY 490

BIBLIOGRAPHY OF BIBLIOGRAPHIC SOURCES . . 533

RÉSUMÉ EN FRANÇAIS 535

REFERAT AUF DEUTSCH 536

INTRODUCTION

There is available at the present time no general manual in English on methods, apparatus, or results in the field of measuring the energy cost of human work, so that experimentation of this kind in the laboratory or in industry is denied to all except those who have had extensive training in physiology. Since the factors influencing metabolism are so numerous and complex and the sources of error in metabolic experiments are so varied and often unexpected, perhaps it is as well that there is no such manual at hand, for it would be likely to tempt novices into engaging in experimental work before they had acquired an adequate theoretical background.

Our grounds for this fear lie in the fact that the measurement of the energy cost of human work is a most attractive field to the industrial engineer and to the industrial psychologist in that it offers promise of help in the solution to many of their most baffling problems. As a matter of fact, several workers whose training in physiology has been limited have already essayed original investigations using metabolic techniques and have published research which reveals this unfortunate lack of fundamental preparation in physiological theory.

The investigators responsible for this type of work have made reasonable efforts to place their research above criticism and are not to be censured for their attempts to introduce these methods into their attack upon industrial problems. Without these techniques,

industrial psychologists and engineers have been forced to satisfy themselves with conjectural answers to many problems relating to the interpretation of different types of work curves and the ultimate effect of various factors upon industrial efficiency. For example, it has been known that production can be accelerated through certain incentive plans, but it has not been possible to determine whether the acceleration is the result of more efficient modes of work, or the expenditure of excessive energy, or both. Motion studies might help reveal whether the act was being accomplished more efficiently, but there has been no means of detecting the possibility of disproportionately increased energy consumption until it made itself known months later in the form of sickness, accidents, labor turnover, or industrial unrest.

There is a need, then, for a means of rendering accessible to the industrial engineer and the industrial psychologist the literature which will enable them to utilize these methods properly in attacking the broad class of industrial problems indicated, and to enable them to perform fundamental research in establishing the limits and the validity of the methods. It is the writer's object in preparing this guide to the literature on respiratory exchange determination to supply information which will be helpful to such a program in two separate, although related, directions. In the first place, this guide should assist any well-trained psychologist or engineer to plan physiologically suitable, adequately controlled, experimental investigations in the laboratory and in the factory; and, in the second place, it

should supply a basis for critical appraisal of current research reported by workers already employing metabolic techniques. It seems to the writer that these two ends are about equally important, and ultimately equally constructive.

It must be understood at the outset that this monograph consists only of a *guide* to the literature; it is impossible to arrive at any understanding of energy metabolism and respiratory exchange from merely reading this monograph. It is essential that the reader maintain constant access to a good library throughout the period of his study. We have assumed an adequate grounding in scientific method in general on the part of the reader, but we have also needed to assume a complete absence of technical training in physiology.

No investigator can hope to plan experiments in the measurement of energy consumption by respiratory exchange methods without a good working knowledge of the general physiology of digestion, circulation, respiration, and muscular work, and it should be unnecessary to add that eventual practice in the manipulation of the apparatus chosen for the research is equally vital. It would be desirable for the investigator to have direct acquaintance with many types of metabolic apparatus, but familiarity with them through the literature will probably have to suffice for most workers. Judicious selection from the references in this syllabus should enable anyone to give himself a course of training in theory and a general familiarity with methods and apparatus that will be adequate for the purpose.

The bibliography* supporting this study is by no means exhaustive within the subjects covered, and many related fields, such as animal experimentation of all kinds, have been wholly excluded. Those who need further references should consult the bibliographic notes which precede the bibliography. Although this guide was not written for the professional physiologist, the author believes that readers in this group will find the references on methods and apparatus more complete than in any other compilation available in English, and may find certain other sections of interest, particularly certain of the sections on factors affecting metabolism.

*The bibliography covers the literature through 1928 only. The author regrets this delay in publication, but feels that it should be possible for the reader to cover the intervening period without undue difficulty.

I

PHYSIOLOGICAL FOUNDATIONS

Theory of Respiratory Exchange Determination. The theory of metabolic measurement through respiratory exchange determination may be outlined in its bare essentials as follows:

Animals derive the energy they expend in work and heat from the oxidation of foodstuffs within the body. The process involves, among many other things, the withdrawal of oxygen from the inspired air, and the formation of carbon dioxide, which is added to the expired air. A certain minimum rate of such gaseous exchange is a necessary accompaniment of the processes responsible for the maintenance of the body temperature, the beating of the heart, and similar activities, and is an index to what is known technically as the basal metabolic rate. Beyond this point, muscular work involves the oxidation of foodstuffs within the tissues in a fixed proportion to the amount of work done. To determine the amount of work done, then, would be a simple matter if we could in some way discover the amount of food stuff which has been oxidized during the work period in excess of that required for the basal metabolic processes. We cannot hope to measure the amount of foodstuffs oxidized in the body by any direct means, but, ignoring certain qualifications which will be more adequately treated in later sections, we can derive a computed figure of the amount if we know either the volume of oxygen which has been consumed or the volume of carbon dioxide given off during the period in question.

Fatigue and Energy Cost. Even this brief outline should serve to save the reader from confusing the concepts of "fatigue" and "energy cost." This is a distinction which has not always been clear, and even today the two concepts are sometimes mingled in a way that indicates rather hazy thinking on the matter. This is recognized, for example, by Strauss and Bandmann (547). In their review of the methods of fatigue measurement they commend the accuracy of metabolic methods but recognize that the concept of energy exchange is not equivalent to the fatigue concept. They suggest that metabolic methods should be particularly useful in studies of the effect of rest pauses and cite in this connection Hill, Long, and Lupton's (285) study of the recovery period. Frois and Caubet (552) have also criticized respiratory exchange measures as a test of industrial fatigue.

Polakov (468, 469), one of the few American workers to utilize respiratory exchange techniques, has fallen into the error just mentioned in his proposal to employ carbon dioxide as an index of fatigue. This proposal was discussed at a meeting of the American Society of Mechanical Engineers (10*a*) and was ably criticized by F. B. Flynn in a written discussion submitted after the meeting. The general tenor of the discussion, however, remained more or less favorable to the method, and Dana (152) and others have quoted Polakov's articles on the subject.

Other workers who have distinguished more carefully between these concepts are such men as Amar (9) in France and Atzler (16, 19, 21) in Germany. The

reader may also be referred to the work of Herbst and Nebuloni (270) in this connection. Waller (602) made wide use of carbon dioxide production as a measure of energy cost, and discovered a progressive increase in cost with continued muscular work. He attributed this decrease in efficiency to the effect of fatigue, but it must be understood that Waller made a clear distinction between the two concepts and was merely interested in discovering their relation to each other.

In an article by Page (458) the energy cost concept is defended as being more susceptible to meaningful quantitative treatment than any of the usual fatigue concepts, and hence of more value in the majority of industrial investigations than the industrial fatigue concept. This paper was developed in part from an earlier contribution by Muscio (443) and in it the stand is taken that there is little relation between the subjective concept of fatigue, the physiological concept, and the industrial output concept, and that the concept of energy cost provides a more generally useful working basis for the industrial psychologist and engineer than any one of these. Because of the similarity of the problems to which fatigue studies and energy cost determinations may be applied, however, we have included a number of the more relevant references to the fatigue literature in our bibliography, and shall refer to them very briefly here.

Literature on Fatigue Measurement. The references to Vernon (569), Florence (194, 196), and Farmer (189) may be taken as examples of output studies of

fatigue, and the papers by Martin (404) and Haggard (241) as expositions in a semi-popular vein of the physiological view of fatigue. More technical treatments on the physiological side are those of Hastings (255) and Durig (178), and a well-organized statement of the problem of industrial fatigue phrased in modern physiological terms has been written by Spaeth (533). This work carries a bibliography of several hundred titles topically classified. Typical physiological fatigue tests are represented by Martin (405), Ryan (493), and Strauss (546), and the entire field of fatigue measurement has been reviewed by Sachsenberg (494) and by Strauss and Bandmann (547). Negative results with physiological fatigue tests are reported by Netschajeff (445) and by Lee and Vanbuskirk (371). A medical approach is represented by Mayers (411) and Ochsner (451).

The work of the Industrial Fatigue Research Board of London is briefly described by Vernon (275) and by Wilson (616). Wilson (615) and Bordas and Courtier (107) have discussed the value of fatigue prevention and Wilson presents a plea for the international pooling of information and facilities through the Geneva office. There seems to have been no action taken on his proposal, however.

General Physiology. Several modern books on elementary physiology are available, of which Mottram (437) and Douglas and Priestley (170) may be taken as examples. Mottram represents a recent popularization of physiology useful in establishing an integrated background. There are chapters on nutrition, respira-

tion, circulation, and muscular work, as well as the other typical physiological headings. More advanced treatments will be found in Stewart (538*a*), in Bayliss (38), and in Howell (299). These are standard text-books of college grade. The reference to Waller (580) is rather old but has been included because of the importance of his later studies on the energy cost of various types of industrial work. In spite of the early date at which the book was written, his chapter on respiration shows strikingly modern insight, and the other chapters on nutrition and related subjects are in many instances subject to but slight revision now. Bainbridge (30) has written a comprehensive book on the physiology of muscular exercise. The entire volume is organized principally as a series of abstracts of the numerous French, German, and English works included in his extensive bibliography. Evans (186) has chapters on recent developments in muscular physiology, muscular contraction, oxygen consumption, carbon dioxide output, and related topics.

Stiles (539) has written a clear and readable introduction to the subject of nutritional physiology. The book may be highly recommended as introductory material for the novice. Lusk (387) is the standard English reference on this subject and is written from a more advanced standpoint. His *Fundamental Basis of Nutrition* (389) is a more recent work on this subject.

It would be futile to demand that the engineer or psychologist acquire a knowledge of biochemistry before starting experimentation, but he should at least familiarize himself with the general scope of this field.

Morse (435) has written a compendious student manual which should prove useful in providing technical background for reading modern writings on the physiology of muscular exercise. On pages 23-26 of this reference will be found periodical and handbook references of a general nature. The book also contains useful chapter bibliographies. Bodansky (92) has written a readable textbook on the subject, although he necessarily assumes a knowledge of organic chemistry on the part of the reader. Pryde (474) and Sumner (550) may be cited as further references in the field of biochemistry.

An attempt to present physiological concepts from the engineering point of view has been made by Briggs (113) in an article which stresses the resemblances between the human body and a steam engine and makes a plea for the study of the human body by the engineer. This study will be referred to later in connection with his concept of the "crest load" or the point which cannot be exceeded without overloading the human machine. Dana's (152) book is also useful for engineers, although the physiological phases are not covered very adequately. McCurdy (418) has written a textbook on the physiology of exercise. This is a good general study with bibliographies at the close of each chapter. The book may be recommended to the general reader. Another book in which the mechanical and engineering aspects are as prominent as the physiological ones is *The Human Motor* by Amar (10). This book treats of the laws of mechanics in a purely physical sense and then applies the laws to the structure and mechanics of the human body. There are chapters on

the chemistry of alimentation, on the mechanics of human energy expenditure, on the factors affecting the economy of human work, on the various types of apparatus used in diverse kinds of human measurement and on the variety of metabolic studies which have been made of walking, climbing, bicycling, and industrial work. McDowall (419) discusses in a non-technical way the physiology of exercise and mental work in their relation to industry.

Basal Metabolism. The reader who has covered even a few of the references given above will realize the importance of understanding basal metabolism and its relation to the metabolism of muscular work. Basal metabolism represents the minimal energy consumption of the body or the energy requirements for maintaining of the body temperature, for providing the energy for the beating of the heart, and for other similar physiological processes of a fundamental and continuous nature. The best-known manual in English is probably that of DuBois (172). This book, written for medical students, contains summaries of the relevant laws of physics and chemistry in their relation to basal metabolism and chapters on nearly all phases of this subject. It may be highly recommended to the beginner on account of its wide scope, profuse references, and scholarly organization. One of the best brief presentations in English of the theory and practice of basal metabolism will be found in the first two chapters of King's *Basal Metabolism* (321). There is a bibliography of about three hundred references, most of which are recent and valuable. Probably the most

complete manual in German is that of Grafe (225), and a useful French manual is that of Terroine and Zunz (558).

Many of the Carnegie Institute of Washington publications carry useful digests and discussions of various phases of basal metabolism. Benedict and Carpenter (70), for example, discuss the significance of various factors in the measurement of basal metabolism, and Harris and Benedict (251) give a more technical account of a biometric study of basal metabolism in man.

Very interesting accounts of the history of basal metabolism, fundamental concepts underlying it, its measurement and the various factors influencing it have also been written by Benedict in 1925 and 1928 (56, 63). The historical point of view is represented also by Lusk (391) and Thannhauser (559). Lusk's history of metabolism runs from Socrates and Hippocrates to Rubner and Zuntz and is a document full of human interest. A very complete summary of the experimental work on basal metabolism was prepared by Boothby and Sandiford (102) in 1924. Their discussion is based upon a bibliography of 697 titles. Talbot (555) has summarized the literature on basal metabolism of children, with a bibliography of 169 titles. Lusk (390, 392) discusses early work on developing the law of surface area. Miscellaneous references of lesser importance are those of Benedict (52), Kahn (314), and Porges (472).

General Metabolic Exchange and Energy Metabolism. The student who is unable to read German will find the majority of reference manuals on this subject

closed books to him. A study of the appropriate sections of the references quoted under the heading of basal metabolism, however, may make it possible to dispense with the assistance of the more general manuals for most practical purposes. The leading manuals and texts are listed in chronological order below.

The textbook by Tigerstedt (560) will be found to be quite rich in references earlier than 1905. The reference to Magnus-Levy (397) is available in English translation and hence may be read by the general reader interested in the early development of the physiology of metabolism. Johannson (308) contributes pictorial illustrations of the various types of respiration apparatus available in 1910. Lefevre (373) has written a very broad and inclusive treatment of the whole field of metabolism, direct and indirect calorimetry, heat regulation and other related physiological topics. It is still of considerable value as a reference. The article by Murlin (440) carries descriptions of all the standard methods and a good discussion of theory and practice. The most serious limitation of his contribution is that there are no specific references, although the article is filled with quotations from many workers. Grafe's (225) manual treats all phases of general metabolism in a comprehensive way and with adequate reference to the literature. It may be highly recommended to those to whom it is available. The Abderhalden (3) handbook is not so useful for our purposes as its title might imply. Only one article in the collection is of interest in studies of respiratory metabolism during work, and this one, Johannson

(309) is not primarily concerned with gaseous exchange, but considers it only in its relation to general metabolism. The review by McCann (412) is of more value to the physician than to workers in industry. There is some discussion, nevertheless, of the cost of work and the effect of miscellaneous factors on metabolism. There is a bibliography of over three hundred titles. Volume IV, Part 10, of the Abderhalden (2) series consists of an exhaustive treatment of gaseous metabolism and calorimetry in the form of a symposium by about twenty leading authors in these fields. It comprises a complete working manual on all types of apparatus and methods. Atzler's contribution to the symposium *Körper und Arbeit* (18) outlines the general theory of metabolism and its measurement. The reference to Caccuri (118) was not available to the reviewer. One of the most modern manuals of practical scope is that of Knipping and Rona (343). The sections on energy metabolism, pages 98-195, and work metabolism, pages 223-240, constitute a clear description of modern techniques, well organized and illustrated with splendid cuts and diagrams. The best modern methods of both direct and indirect calorimetry are discussed, and the book includes many useful points of practical technique. The textbook by Krauss (348) is another comprehensive treatment of indirect calorimetry by both open and closed circuit methods. Krauss describes and illustrates a wide variety of experimental methods but is not much concerned with work experiments. There are several useful tables and a bibliography, pages 305-312, of over one hundred titles

on gaseous exchange, chronologically arranged from Lavosier to date. In this section should also be mentioned the review by Murlin (442) of metabolism in infancy and childhood. This article treats its subject broadly and is supported by a bibliography of over two hundred titles but is not of much interest in connection with work metabolism in spite of the general introductory material included. The majority of the above references are of a theoretical or general nature; references more specifically concerned with apparatus and methods will be supplied in a later section.

Respiratory Physiology and Blood Chemistry in Relation to Muscular Work. It is essential to know something of the physiology of internal and external respiration and their relation to blood chemistry and pulmonary ventilation respectively in order to understand the technical material on the biochemistry and dynamics of muscle action and the effects of muscular work on the body as a whole. Although the majority of investigators will need only a few specific references on this subject in addition to the material already covered in the general references mentioned in preceding sections, it is desirable for all experimenters to be familiar with a few of the general references and in certain instances more specific material may be needed. It should be understood, then, that all of the references given below are not recommended for general reading.

Newcomers in this field are fortunate in having available Haldane's (243) splendid historical treatment of the scientific investigation of the problem of respiration. This book provides the best possible sub-

stitute for the experience of having worked many years in the field of respiration and the opportunity of following its development discovery by discovery. The volume is not always easy reading, although limited to elementary mathematics and chemistry. There are chapters on such subjects as carbon dioxide and the regulation of breathing, the nervous control of breathing, the blood as a carrier of oxygen, the effects of the want of oxygen, blood reaction and breathing, blood circulation and breathing, and the effects of various factors upon respiration. There is an appendix on blood chemistry.

A. V. Hill is well known not only for his scientific contributions to the theory of muscular exercise, but also for his semi-popular works on muscular activity (281-284). These references will be considered more in detail in a later section and are mentioned here only because the breadth of their scope makes them appropriate. A somewhat earlier reference which remains standard is that of Krogh (354). All phases of the subject of respiratory exchange of animals and men are discussed and there is a well-organized and extensive bibliography to 1916. Reference is made here to the early work of Hanriot and Richet (247) principally as a matter of record and classification. The reference will be considered again in connection with special methods in work experiments.

Gesell (212) has written a review of the current theories of the chemical regulation of respiration, with emphasis on his own view. The article is concise and well written and has a bibliography of 141 titles. He wrote

a similar article the next year in *Science* (211), expressed in more popular terminology. Scott (509) may be taken as an example of technical research of the type reviewed by Gesell. Scott's experiments seemed to demonstrate that "undissociated carbon dioxide acts as a specific respiratory hormone. Therefore the physiological effects of carbon dioxide on respiration cannot be attributed solely to its acid properties when in solution" (opposed to the view that carbon dioxide excites the respiratory center only through its action on the hydrogen ion concentration of the blood, to which the body makes adjustment). See also Bald (31) for a discussion of the regulation of respiration in terms of blood chemistry.

The first five chapters of Henderson and Haggard (266) constitute a useful exposition of modern respiratory physiology. This may be recommended to all workers. DuBois-Reymond (99) contributes an article on methods of studying the mechanics of respiration. Takahira (553) is quoted by DuBois in *Basal Metabolism,* page 21, as a good general review of the subject of respiratory metabolism. This was not available in September, 1928, in the John Crerar Library, the University of Chicago Library, or the Library of Congress, Washington. The reference to Brunton (115) consists of a collection of clinical and experimental papers of a minor nature, most of which were originally published in the eighties. Campbell, Douglas, and Hobson (121), in discussing their experiment on the respiratory exchange of man during and after muscular exercise, presented an extended account of

the physiology of the various processes involved. They were concerned particularly with the rise in the respiratory quotient immediately following work and the reasons for various changes in gaseous metabolism. This reference will be discussed later in connection with muscular work. In the article by Krogh and Lindhard (356) will be found a theoretical discussion of the mechanism which provides the very rapid adaptation of the respiratory and circulatory systems to sudden muscular exertions.

Aitken and Clark-Kennedy (6) divided a single expired breath into six successive portions in their study of the fluctuations in the composition of alveolar air during the respiratory cycle in muscular exercise. A preliminary announcement of their apparatus, methods, and conclusions was written in 1927 (5) in which it was stated that "during the tenth minute of moderate work on a bicycle ergometer a single breath, two to three and one-half liters, is divided up into six successive portions by means of a special apparatus. The carbon dioxide concentrations in the six portions are plotted against the respective volumes, a smooth curve being drawn. . ." They found a typical S-shaped curve from the origin to about 1100 cc., followed by a straight portion sloping gently upward. The average carbon dioxide percentage is lower than that of the last 10 cc. alone. Briggs (112) develops his conception of physiological overload in some detail in a study on physical exertion, fitness, and breathing. "When exertion of steady, increasing magnitude is undertaken the expired carbon dioxide percentage first rises and then falls."

He considers that this marks the point beyond which there is physiological overload.

Of practical importance in the selection of measurement techniques is the assurance from the experimental results of Periera (463) that oxygen consumption when one is breathing pure oxygen differs practically not at all from the consumption when breathing atmospheric air. The validity of this assertion has been attested to by a large number of investigators. Padget (456) has shown that inspiration of carbon dioxide causes an immediate rise in the carbon dioxide tension of arterial blood, but a lag in other responses, perhaps due to time required to saturate the tissues, especially the respiratory center. The reaction of any one individual to carbon dioxide is extremely constant but it varies greatly among different individuals. Another experiment on the effect of breathing different concentrations of carbon dioxide is reported by Goldstein and DuBois (221). Means (422) has written an extensive review of the many phases of the subject of dyspnoea. His bibliography and treatment of general subjects may be of some interest to investigators in work metabolism. Perwitzschky (464) has investigated the temperature and moisture of air in air passages under normal environmental conditions and normal depth of breathing in resting man. Jahn (304) has studied the relation between oxygen consumption and carbon dioxide concentration in an experiment on the specific dynamic action of food and the laws of gaseous exchange. Douglas and Haldane (168) studied the capacity of the air passages under varying physiological conditions. Their

data on the results of experiments on walking, given originally in tabular form, were plotted by Page (458), who was interested in demonstrating the sensitivity of *ventilation rate* alone as an index of energy metabolism.

Ponzo (470, 471) discusses psychological influences upon the rate and character of breathing. There is momentary suspension of breathing during close attention, acceleration and retardation accompanying thoughts with different affective tones, and disturbance due to slight laryngeal movements during thinking. Specific examples are given, such as slower respiration with work at scientific instruments and during reading and other mental work.

Mathieu and Schaeffer (409) report briefly on experiments indicating an inverse relation between carbon dioxide concentration and respiratory frequency. It was possible to establish certain relations of a definite sort when the averages of groups were taken. Of considerable importance in connection with the Waller method of estimating energy consumption (to be discussed in a later section) is the article by King and Cross (323) on superventilation and carbon dioxide elimination. The reference will be discussed in further detail in connection with the theoretical definitions of the Waller method. Durig (179) reports experiments on gaseous metabolism performed in connection with the Mount Rosa expedition. Useful theoretical discussions of respiratory exchange and respiratory physiology will be found in several of Benedict's studies of which Numbers 54, 58, 70, and 82 (of our bibliography) may be taken as examples.

It is difficult, and perhaps not highly essential, to separate the discussion of blood chemistry in relation to muscular work from the consideration of respiratory physiology in the same connection. The two topics have necessarily been treated together in most of the references given above so that the student should now understand the relationship of respiratory measurement to the fluctuations in the composition and hydrogen ion concentration of the blood stream. A good deal of modern work is being done on direct measurements in blood chemistry rather than the measurement of respiratory exchange. For this reason we shall need to mention a few references specifically related to blood chemistry, even though our principal interest in preparing this guide to the literature is in connection with indirect calorimetry by means of respiratory exchange measurement.

The articles by Hastings (254, 255) are useful studies of changes in the blood following muscular work. His "Physiology of Fatigue" (255) constitutes a significant contribution to the theory of muscular exercise. It is based upon a study of the hydrogen ion concentration and the composition of the blood of dogs working on a tread-mill. The articles by Barr, Himwich, and Green (34) are written in three parts: Part I, "The Changes in the Acid-Base Equilibrium Following Short Periods of Vigorous Muscular Exercise"; Part II, "A Comparison of Arterial and Venous Blood Following Vigorous Exercise"; Part III, "The Development and Duration of Changes in the Acid-Base Equilibrium." There are good bibliographies with each section. The

contributions of Lundsgaard and Möller (386) are also in three parts: Part I, "Oxygen and Carbon Dioxide Content of Blood Drawn from the Cubital Vein Before and After Exercise"; Part II, "Oxygen and Carbon Dioxide Content of Blood Drawn from a Cubital Vein at Different Intervals After Exercise"; Part III, "Effect of Varying the Amount and Kind of Exercise." In the paper by Bock, Dill, Hurxthal, Lawrence, Coolidge, Dailey, and Henderson (88) will be found a description of the principal physio-chemical properties of the blood of a normal man in a steady state of work and the changes in the blood accompanying the change from rest to work.

Van Slyke (567) has written a technical review of the chemistry of the carbon dioxide carriers of the blood, in terms of buffers, hydrogen ion concentration, etc. Of theoretical interest principally is the article by Schneider and Truesdell (506). An increase in the carbon dioxide content of the blood in man was found to raise the blood pressure and increase both the volume and frequency of breathing. In their studies several types of physiological changes were noted in detail. Barcroft (33) has written a comprehensive manual on *The Respiratory Function of the Blood*.

The attention of psychologists who may be using these bibliographic notes is drawn to the article by Starr (537) on the relationship of high alveolar carbon dioxide tension to the etiology of stammering. This subject is not of interest perhaps to industrial engineers, but is an interesting application of the study of blood

chemistry to a practical problem often studied by psychologists.

The Biochemistry and Dynamics of Muscle Action. Most of the work on this subject has centered about the investigations of the British physiologist, A. V. Hill, and his colleagues. He has contributed extensively both to the literature of pure science and to semi-popular exposition. Early articles of a reasonably non-technical nature were written in 1924 (282) and 1925 (283). This latter was written with characteristic clarity and conciseness and may be highly recommended to those wishing to read only twenty-five pages of theory. In 1926 (281) he published a more extensive treatment of muscle physiology with ample citations of the literature and descriptions of recent work. It is written in as non-chemical and as non-mathematical a way as it is possible to write such a book. A more widely-known account is his *Muscular Movement in Man* (284), published in 1927. The first half of this book treats of the modern physiology of muscular work and the last half is concerned with the viscosity of human muscle, the dynamics of sprinting, the mechanical efficiency of human muscle and many matters of theoretic interest to the biological chemist and to scientifically inclined athletes. The bibliography contained in his book, *Muscular Activity,* is carried up to date, i.e., from 1924 to 1927.

Hill and his collaborators have produced an extensive scientific literature which can be only briefly sampled here. Hill and Lupton (286) present an extended and semi-technical account of the physiology of

exercise in man. This work is not a report of an experiment but is a general discussion with experimental illustrations. The reference is valuable for theory and bibliography and constitutes a more complete treatment of some of the material found in *Muscular Movement in Man*. The reference to Hill, Long, and Lupton (285) is one of an extensive series in the *Proceedings of the Royal Society*. Volume 96 contains Parts I to III, which consist of an extended treatment of muscle physiology with bibliography and mathematical and chemical discussion. "Lactic acid in muscle does not, to any serious extent, directly turn out carbon dioxide from bicarbonate. It combines with sodium protein and raises the hydrogen ion concentration; the elimination of carbon dioxide which results is the consequence of the induced activity of the respiratory system." The respiratory quotient was found to fluctuate widely after severe exercise but was not seriously affected by mild exercise. It was found to depend upon lactic acid, with its influence upon the respiratory center through the effect of hydrogen ion concentration. The recovery process is studied in detail in Parts V and VI. Careful distinctions are drawn between the effects of "severe" and "moderate" exercise, and their physiology is separately discussed.

Hill's theory that muscles utilize carbohydrate only during work is the center of active controversy. Hill's view is supported by Bock, vanCaulaert, Dill, Fölling, and Hurxthal (91) who find that glycogen is the principal immediate source of energy for muscular contraction. Krogh and Lindhard (357) publish experi-

mental results indicating a higher mechanical efficiency under carbohydrate fuel, indicating that it is carbohydrate that is actually utilized in the reactions and that fat must first be converted into carbohydrate, with consequent loss of mechanical efficiency. DuBois (174) also supports Hill's views in a study of metabolism in disease and in health. He emphasizes the difference between the nature of the diet and the nature of the foodstuffs actually metabolized and shows the necessity of interpreting experimental results in the light of this fact.

Himwich and Castle (292) and Himwich and Rose (293) have studied the respiratory quotients both of resting and of exercising muscle and found that in each case the muscle had a respiratory quotient, not of unity, but of practically the same order as that of the body as a whole. This indicates that not only carbohydrate but also some fat and protein foodstuffs are also consumed. Fenn (190) found the respiratory quotient of frog muscle to be less than unity both while at rest and during exercise. This supports the findings of Himwich and Castle and Himwich and Rose. Rapport and Ralli (477) are also in agreement with these findings. The following is quoted from *Physiological Abstracts,* 1928, No. 1010: "It is suggested by the evidence that in mild exercise of short duration carbohydrate is not the sole source of energy and that fat is utilized, not to replenish carbohydrate stores, but by oxidation to supply energy for muscular exercise. The muscles oxidize usually a mixture of fat and carbohydrate dependent upon what foodstuffs they receive, and their propor-

tion." Another recent contribution is that of Lindhard (382), whose article is abstracted in *Physiological Abstracts,* 1928, No. 2181 as follows: "A re-investigation of the problem with precautions against interference with respiratory movements. Work was performed on a bicycle ergometer at what has been found to be the optimum speed. The experimental results of A. V. Hill and his colleagues on this subject are subjected to criticism, and the author finds no support from his own experiments for the conclusion that muscular work of short duration is done entirely at the expense of carbohydrate." The subject is discussed by DuBois (172), page 46, who quotes Lusk (388) in opposition to Hill's theory that carbohydrate alone furnishes the energy in long-continued exercise.

The dynamics of muscle action are treated by Hill in several of his contributions, both popular and technical. Amar (10) also devotes one or two chapters to this subject. More technical articles are those of Furusawa, Hill, and Parkinson (204) and Hopkins (297).

Cathcart (130) traces the reversal of the modern theory of protein metabolism from the older conception (i.e., that it is the source of energy for muscular work). He shows that protein is concerned in muscular work nevertheless. There are about eighty references in the bibliography, giving the author and periodical but not the title.

A very popular article on the elementary biochemistry of muscle action, although expressed in modern terminology, is the article "Fatigue and Rest" (404), contributed by Martin to *Industrial Psychology.*

The Physiology of Muscular Work. There is, of course, considerable difficulty in separating the discussion of this topic from the discussion of the biochemistry and dynamics of muscle action, treated in the preceding section. Some of the more general writers such as Hill (281-284) and Bainbridge (30) may be referred to just as appropriately here as they were above. It must be obvious that these two sections deal merely with two aspects of one fundamental problem and that discussions of the one will usually be phrased in terms of the other. Atzler (20) has written a concise review of the physiology of muscular work, with about fifty citations to the literature. Grafe (225) also has a useful chapter on this subject with 69 titles in his bibliography. The article by Riesser (483) appears to be a general treatment of this subject, but was not available to the reviewer. The Vienna Letter (578) in the *Journal of the American Medical Association* is a very sketchy report of Professor Rubner's lecture on "Modern Conceptions of the Physiology of Work." Magne (396) has written a good general account of respiratory changes during muscular exercise, with considerable tabular data quoted from other experimenters. Some of the contributors to *Körper und Arbeit* (19) have written articles of a general nature. Herbst (269) and Mangold (400) may be cited as examples.

Some of the most thorough investigations of the physiology of muscular exercise have been carried on by Benedict and his colleagues at the Carnegie Institution of Washington. The work of Benedict and Carpenter (69) on the influence of muscular and mental work on

metabolism is classic in this field and carries a good summary of earlier work on mental and muscular metabolism. The work of Smith (525) is another typically thorough Nutrition Laboratory study. Smith found that physiological adaptation to work occurs in large part within 30 seconds and is complete within 3 minutes. Recovery was found to be not quite so prompt. Henderson, Dill, vanCaulaert, Fölling, and Coolidge (262) have written a short article pointing out that the results of many experiments show convincingly that *there is an analogy between the behavior of the human* mechanism and that of a machine which adjusts itself to a steady state of work no less smoothly under a heavy load than when idling. It is possible for the human machine to carry on smoothly while arterial blood, and therefore cell environment, remains approximately constant during heavy work. Other articles of a general nature are the contributions to the symposium in the International Labor Review (552) and that of Knoll (344) who studied internal and external respiration in certain of the major sports.

Two or three studies on the relationship existing between physical and mental work may be mentioned here. There is, of course, the study of Benedict and Carpenter quoted above, and there are several more recent studies of the energy cost of mental work. It is appropriate here to refer only to the studies of Day (157), Gillespie (215), and McDowall (419), all of whom were concerned with the influence of both physical and mental work on such processes as breathing and circulation. These references will be discussed in

greater detail in the section devoted to mental work.

Early studies in the physiology of exercise utilized rather simple physiological measures, including respiratory exchange determinations such as we are advocating for widespread use in industry. Investigators working upon theoretical problems, however, have developed methods and apparatus of greater and greater refinement for various technical purposes. Higley and Bowen (278) describe several different early methods, including their own (published in 1905). Krogh (349) discusses the accuracy of respiratory exchange determinations in experiments of very short duration. Kaup and Grosse (317) discuss older methods and make new suggestions. Boigey (97) illustrates his graphical methods of determining respiratory exchange during work. Further discussion of respiratory exchange determination methods and apparatus will be reserved for a later section. Schneider (500) proposes a cardiovascular rating as a measure of physical fatigue and efficiency. Since this rating depends upon the effect of exercise on pulse rate and blood pressure, it seems pertinent to mention it here. Henderson and Prince (267) discuss the oxygen pulse and the systolic discharge. The oxygen pulse is defined and its use explained in connection with the physiology of exercise.

The technical developments mentioned in the above paragraph are well illustrated by techniques which have recently been made available for determining the circulation rate in man. Lindhard (380) uses the nitrous oxide method in his study of circulation after cessation of work. Bock, Dill, and Talbott (89), in

recent work from the Fatigue Laboratory, Morgan Hall, Harvard University, and the Medical Laboratories of the Massachusetts General Hospital, prefer the so-called "Haldane" method to either the nitrous oxide method or the ethyl iodide method. In the "Haldane" method the circulation rate is determined indirectly from the alveolar carbon dioxide tension. Bock, vanCaulaert, Dill, Fölling, and Hurxthal (90) continue the study quoted in the preceding sentence. They determined blood flow, pulse, blood pressure, lactic acid in the blood, carbon dioxide dissociation curves and Haldane analysis of expired air in a study of dynamical changes occurring in man at work. There is a bibliography of 28 references. Dill, Hurxthal, vanCaulaert, Fölling, and Bock (161) demonstrate that the automatic sampling device used in the determination of the rate of blood flow by the ethyl iodide method gives values much too low for the carbon dioxide pressure of arterial blood, and Dill, Lawrence, Hurxthal, and Bock (162) show that Haldane-Priestley samples of alveolar air collected during exercise at the beginning of expiration measure approximately the average carbon dioxide pressure of arterial blood. Another careful experiment on hydrogen ion concentration of the blood and alveolar carbon dioxide tension is that of Arborelius and Liljestrand (14). Hough (299) describes an improvement in the Haldane method of collecting samples of alveolar air. This is in an early study of the physiology of muscular exercise. Goiffon (218) describes a recent device intended to simplify and standardize the taking of samples of alveolar air. A dis-

cussion of the determination of the circulation rate in man at rest and in work has been written by Boothby (100).

The technical studies of the physiological adjustment of the human body to muscular work are entirely too numerous to permit more than a casual sampling here. Cook and Pembrey (144) and MacKeith, Pembrey, *et al.* (395) have conducted elaborate studies of the alveolar carbon dioxide tension, acidity of urine, circulation rate, and similar factors in heavy exercise in which "second wind" is ordinarily experienced. Cook and Pembrey report that "second wind appears to be an adjustment of the circulatory and respiratory systems to the demands of the muscles for an adequate supply of blood. Carbon dioxide is the chief factor in affecting the accommodation." In the latter article by MacKeith, Pembrey, *et al.* this idea is amplified and the adjustment explained principally in terms of disturbance and re-establishment of the acid-base equilibrium of the body. Similar studies of adjustment of the human body to muscular work are those of Talbott, Fölling, Henderson, Dill, Edwards, and Berggren (556), Barr, Himwich, and Green (34), Lundsgaard and Möller (386), Herxheimer and Kost (272), and Gordon (222). Other technical studies of interest here are those of Martin and Gruber (406), Douglas (166), and Long (385).

Parts I and III of the series by Simonson (519, 521) are similar in nature to the studies listed above. Part IX of the series by Viale (575, 577) may also be taken as an example of this type of investigation. Other

special studies on the correlation between respiration, circulation, and oxygen consumption during muscular work are those of Mobitz (427), Ledent (369), Cassinis (128), and Boigey (27).

Studies in which the principal emphasis was placed upon the recovery period following muscular work are those of Sargent (497), Campbell, Douglas, and Hobson (121), Condero (142), Herxheimer, Wissing, and Wolff (273), and Liebenow (376). Sargent found that recovery was extremely rapid, particularly for the first ten minutes after exercise. The rate was found to vary with the subject and the severity and duration of exercise, however, so that total oxygen consumption during recovery cannot be estimated by applying a fixed correction to observed partial recovery. An extended and thorough-going account of the physiology concerned and especially of the rise in the respiratory quotient will be found in the reference by Campbell, Douglas, and Hobson.

Studies having as an important aim the determination of the nature of the foodstuffs concerned in muscular work are those of Furusawa (201), Henderson and Haggard (265), and Marsh (402). These studies will be considered again in connection with our discussion of the respiratory quotient and so will not be abstracted further here. The work of Curtis (151) may be taken as an example of a study of the influence of the thyroid gland on working metabolism. Simonson (518) has studied the effect of forced breathing on recovery from muscular work.

The following studies may be mentioned as being of

interest in a practical way as well as in their theoretical contributions: Benedict and Cathcart (72), a study of the efficiency of the human body as a machine; Atzler, Herbst, Lehmann, and Müller (24), studies of various methods and postures used in lifting weights; Atzler, Betke, Lehmann, Sachsenberg, *et al.* (21), studies on work and fatigue; Atzler (19), mentioned previously as a useful symposium; Hansen (248), a determination of optimum speed of work on a bicycle ergometer; Hill and Campbell (289), a study of the effect of atmospheric cooling upon efficiency during work, and Walther (605), on the "techno-psychology" of work (not available to the reviewer).

The Respiratory Quotient. The respiratory quotient has already been discussed in connection with Hill's view that carbohydrate is the source of energy in muscular work. Several studies supporting this view were cited as well as a number in opposition to it. It is essential now to consider the matter in more detail as it is necessary to understand the use of this ratio as an index of the energy value of the oxygen consumed in work experiments. These matters are amply discussed in a number of manuals listed in a preceding section, for example, Chapter IV of King (321) and several chapters in DuBois (172). Other general discussions of the significance of the respiratory quotient during and following muscular work are found in Amar (7) and in Campbell, Douglas, and Hobson (121).

The most important recent developments have been in the direction of demonstrating the great complexity of the factors which interact to produce the respiratory

quotient as determined at any instant. Cathcart and Markowitz (135) show that the ratio existing between the volumes of carbon dioxide and oxygen represent, not a single, relatively simple physiological phenomenon, but that it is the sum of an infinitely large, relatively unknown series of phenomena. Conybeare and Pembrey (143) also contend that the respiratory quotient is a resultant of many factors and that the older view that it is composed only of varying contributions of the three classes of foods (R.Q. 0.7 for fat, R.Q. 0.8 for protein, and R.Q. 1.0 for carbohydrate) is no longer tenable. The component above R.Q. 1.0 can be explained by the conversion of carbohydrate into fat and the component below 0.7 can be explained by the conversion of fat into carbohydrate. Fries (198) finds that the respiratory quotient fluctuates widely throughout the day and that the average of a few short period tests does not always represent the daily respiratory quotient. Furusawa (201) showed that it is impossible to determine from the respiratory quotient what substance is being oxidized in the muscle itself during exercise. Hendry, Carpenter, and Emmes (268) declare that when basal metabolism only is desired it is not essential to determine the respiratory quotient but that it must never be neglected when work of a scientific nature is being done. Knipping (331), on the other hand, insists upon the necessity for determining the respiratory quotient in routine basal metabolism tests as well as in all other experimental investigations.

Benedict, Emmes, and Riche have been quoted by DuBois (172) as showing an R.Q. of .88 after a meal

rich in carbohydrate as compared with .82 as a normal basal. Many experiments since that date have indicated an effect of the preceding diet on the respiratory quotient even after digestion has ceased, and Marsh (403) in a recent study showed that "the respiratory quotient for the excess metabolism in moderate work is more constant if the diet is controlled than otherwise. On mixed diet it is .95, on carbohydrate it rises and on fat diet it is .83-.80. The net efficiency of a subject first on mixed, then on carbohydrate and lastly on fat diet decreased during the fat diet slowly up to the eleventh day, when it fell markedly." The above is quoted from *Physiological Abstracts,* Volume 12, No. 3890.

A short article by DuBois (173) is available, showing graphically the respiratory quotient and the percentage of calories furnished by protein, fat, and carbohydrate. Smart (524) reproduces a slide rule for the calculation of the respiratory quotient and supplies the mathematics of its construction. Further details of the use of the respiratory quotient in determining energy consumption will be given in the next section.

Hill's contention that glycogen supplies the energy of muscular work has already been discussed in another connection but should be more amply treated here, through the citation of the more significant studies in a more or less chronological order. Cook and Pembrey (144) found the respiratory quotients of their men at rest ranged from .75 to 1.03, with a mean at .90, and that after exercise they ranged from .81 to 1.37, with a mean of 1.00. This study, which would appear to

support Hill's views in part, was followed by a similar investigation by MacKeith, Pembrey, *et al.* (395) about ten years later. Krogh and Lindhard (356) found that the respiratory quotient rose rapidly to or above unity at the beginning of heavy work, and the same authors (357) found a higher mechanical efficiency under carbohydrate fuel. Krogh (351) has written a short summary of this later study. DuBois (174) also supports Hill's views, and in the very recent work of Bock, van-Caulaert, Dill, Fölling, and Hurxthal (91) the respiratory quotient was found to be fairly constant at about .95. This would indicate that glycogen is the principal immediate source of energy of muscular contraction.

The studies of Himwich and Castle (292) and Himwich and Rose (293) have already been mentioned as being contradictory to Hill's view. In the same discussion mention was also made of the work of Fenn (190), Rapport and Ralli (477), Lindhard (382), and Lusk (388). To these studies opposing carbohydrate as the sole source of muscular energy should be added the early study of Morgulis (434) who believed that respiratory quotients of 1.00 and over obtained during muscular exercise should be interpreted as indications of faulty technique rather than indicating the use of pure carbohydrate of the transformation of glycogen into fat, as would be implied for a respiratory quotient over 1.00. He accepts Zuntz's work on the respiratory quotient in which the respiratory quotient was found to be unaffected by muscular activity. Henderson and Haggard (265) also report of their experiment "that

the most significant result of these observations is the conclusive evidence which they afford that in whatever proportion fat and sugar are being burned during rest just before the exercise, they are burned in nearly the same proportion to produce the energy for doing work or for the recovery process in the muscles." Marsh (402) also supports Zuntz's view that both fat and carbohydrate are used for energy in muscular work. A study of the mechanical efficiency of the body on carbohydrate, fat, and mixed diets was made by Severinghaus, Reynolds, and Stark (514), who found that the net efficiency was the same on each diet, although there was ketosis under the fat diet and fatigue was produced during work. This work may be taken as indirect evidence that the body may use varying proportions of these fuels indiscriminately without the necessity of conversion to carbohydrate. Furthermore, Wilson, Levine, Rivkin, and Berliner (617) found that the respiratory quotient of the extra metabolism due to exercise was uniformly less than unity. The reference to Van Slyke (566) on *The Relation of Carbon Dioxide and Oxygen* is of significance principally to physicians and is mentioned in discussing the respiratory quotient only because of the apparent pertinence of the title.

Calorie Computation. The reader who has achieved a good working knowledge of the physiology of muscular exercise as treated in the various references given above should now be in a position to consider the problem of computing the energy exchange of an experimental period in terms of calories consumed. We have considered the significance of the respiratory quotient

and various other factors entering into the computation of energy exchange. There still remain the subjects of surface area and certain other factors playing a part in these computations, but these matters may be more appropriately deferred to a later section, since a superficial understanding of them is sufficient for the purpose at hand.

The general manuals cited in the section on *Basal Metabolism* may be referred to for guidance in computing the energy exchange from experimental data. DuBois (172) and Grafe (225) may be cited again in this connection. Sanborn (496) also treats this topic briefly. Janet (305) contributes a general discussion of computation and supplies certain useful tables, nomograms, and formulae. His height-weight surface area nomogram has been reproduced by Wardlaw (607*a*). Amar (10) describes the computations in work experiments briefly.

The most complete treatments of this subject are found in articles primarily concerned with some of the open circuit methods of determining respiratory exchange. The theoretical and practical differences between open circuit and closed circuit methods will be treated in a later section as the computation methods involved may be understood and followed without technical knowledge of the distinctions between these two broad lines of experimental approach. Henderson (264) appends a good set of student laboratory instructions to his description of respiratory exchange determination by means of typical open circuit methods. Cathcart (131) has written an elementary statement of

the theory of indirect calorimetry which would be considered somewhat crude and out of date at the present time, but which contains condensed instructions concerning the methods of computation and carries certain useful conversion tables as well. Dautreband and Davies (155), Labbé and Stévenin (359), and Boothby and Sandiford (104) also describe open circuit methods and illustrate the computations involved. Klein and Steuber (325), in their discussion of various chemical absorption methods for determining oxygen and carbon dioxide, treat methods of calculation in detail and supply several useful tables. The reference to Wheeler (611) on *Measuring the Energy Cost of Work* is disappointing in this connection as it consists of an inaccurate and misleading description of the Douglas Bag method with detailed quotations from two studies by Langworthy and Barott, with no acknowledgment made to the original investigators.

The computations of energy expenditure from results obtained by means of closed circuit apparatus have been treated by Benedict and Tompkins (82) and others who will be mentioned when we consider closed circuit methods in greater detail.

Stoner (543, 544) describes a simplified data blank and simplified calculations for use with open circuit apparatus. Newcomer (447) has prepared tables and charts by which the basal metabolic rate may be calculated by simply adding five numbers.

Several types of nomographic charts are available. Boothby and Sandiford (106) discuss the mathematics of constructing such charts and publish a variety of

them, as well as correction and reduction tables and a sample of the Mayo metabolic blank. "A series of nomographic charts is described by which the calculation of basal metabolic rate by the gasometer method can be made graphically in less than five minutes without the use of logarithms." Smith and Smith (527) also publish graphs for use in determining basal metabolic rates. Dill and Fölling (160) say of their nomographic charts that "one joins with a thread the percentage of carbon dioxide and of oxygen found in the expired air. The respiratory quotient and percentage of oxygen used can be read directly with an error of less than one in five hundred." Kommerell (345) also publishes nomographic charts for use with the Douglas Bag method.

Hollingsworth (295) makes a plea for wider use of the slide rule by physicians and describes a method of determining basal metabolic rate on a circular slide rule. Smart (524) also reproduces a slide rule for the calculation of respiratory quotients and supplies the mathematics of its construction.

Useful tables to be used in the computations have been supplied by a number of authors, some of whom have already been mentioned. Haldane (242) supplies conversion tables of various kinds; Knipping and Kowitz (342) supply several useful tables, including one giving the logarithms for reduction to 0° C and 760 mm. pressure; Gauss (208) publishes a table combining the corrections for barometric pressure, room temperature, brass scale expansion, and vapor tension; Carpenter (125) assembles tables for reduction to 0° C and 760

mm. pressure, for estimating body surface, the latest basal standards, and factors for converting various units of energy from the one to the other. Roth (485) and Krauss (348) also publish tables which are useful in metabolic rate determination.

Other charts and tables of some interest here are those of DuBois (173) and the Royal Society of London (492). DuBois says of his chart that "almost all the phenomena of respiratory metabolism can be represented on this metabolic map and we can follow the changes which result from the ingestion of protein, fat or carbohydrate." The material supplied by the Royal Society of London is concerned particularly with food requirements under various conditions in terms of calories.

One other reference of value here is that of Gephart, DuBois, and Lusk (210) who contend that in metabolic work the analytical error is seldom much less than one per cent and, since a variation of one per cent is of little significance, it is unnecessary to publish more than three significant figures in the data of metabolism experiments.

Factors Affecting Metabolism. We have already seen that the total metabolism consists essentially of the individual's basal metabolism plus the metabolism of the muscular exercise in which he engages. If there were no other factors than these influencing the metabolic rate it would be a relatively simple matter to determine the energy requirement for any specific type of muscular activity by determining the total metabolism while the individual is engaged in this activity and de-

ducting from that total the basal metabolism, previously determined. Actually, however, the metabolic rate is influenced by a large number of factors both physiological and environmental, so that it is impossible to conduct significant experimentation in work metabolism without knowledge of the possible influence of these factors in order that they may be either controlled or observed in work experiments. It is impossible to classify all of these factors in metabolism in any rigid sort of way, but as a matter of convenience we shall group them under three principal heads. The first group will consist of those somatic and intra-organic factors which are usually subject to observation but which are beyond the control of the experimenter. The second group are somatic and intra-organic factors usually subject both to observation and control, and the third group will comprise those factors which may be thought of as environmental.

The most important of the first group of somatic factors, whether one is computing basal metabolic rates or is engaged in studies of working metabolism, is that of the weight and surface area of the body. This has already been mentioned in connection with calorie computation but will be treated in more detail here.

In addition to the information in the various manuals which have been repeatedly referred to, we may cite the recent comparison which Stoner (545) has made of the various formulae. The Dryer standards are recommended but the point is made in his article that the choice of standards in the determination of basal metabolic rates is coming to be of decreasing

significance since such determinations are now made for purposes of therapeusis rather than diagnosis. The DuBois formula (175) seems to be rather generally accepted and tables of values of the DuBois surface area formula have been worked out by Stoner (542) in a convenient form for reference. Bradfield (111), as abstracted in *Physiological Abstracts,* Volume 13, No. 603, finds that "the average basal metabolism for women is six per cent below that predicted by all standards and is the same as that predicted by Krogh. Krogh's modification of the Aub-DuBois standards gives correct results for women. If the DuBois height-weight formula is used a correction of plus two per cent should be made."

Other general discussions and summaries of the work on surface area have been written by McCann (412), Murlin (441), Krogh (350), Hedon (258), and Boothby and Sandiford (103). Boothby and Sandiford (105) have also presented metabolism data on the subjects going through their laboratory over a period of about five years. They concluded that the DuBois standards were the best available at that time.

Those interested in early controversies concerning the effect of body surface and weight upon heat production may note that Benedict and Smith (81) wrote in 1915 that "it would thus appear that the increase in the metabolism noted with athletes points strongly towards the earlier conception that the catabolism of the body is proportional not to the surface of the body but to the active mass of protoplasmic tissue." The question of the influence of the active protoplasmic mass has

been debated for many years, but Corlette (145), as well as most modern workers, finds that the active protoplasmic tissue is not the controlling factor in basal metabolism. Kaup and Grosse (315) have written a careful survey on the literature on surface area in which the American literature on the subject is criticized as being based upon cases with too little homogeneity. They report their own study on subjects of uniform physical characteristics and they minimize the effect of body size and weight which they regard as minor variables. They find in Darwin and Lamarck a basis for a broader treatment of the subject in terms of the preservation of species.

Sex and age are other somatic factors beyond the control of the experimenter. Benedict (61) has written a general article on age and basal metabolism, and Murlin (442) and Talbot (555) have written general reviews of metabolism in infancy and childhood. Göttsche (223) finds a specific reaction in puberty, and Fleming (192) suggests that the high basal metabolism of the growing child may be partly accounted for by the energy expended in the manufacture of new tissue. Investigators in industry are not likely to find age a highly important variable among the subjects ordinarily dealt with.

Another non-controllable organic factor which has attracted considerable attention in recent years is that of the influence of race. These studies have related principally to racial differences in basal metabolism rather than being concerned with possible differences in the energy cost of equivalent tasks performed by per-

sons of different races. In general, it appears that the basal metabolic rate of northern groups is in excess of that of residents of more tropical climates. The differences are not large and may be obscured by the relatively wider range of individual differences. Recent experimental determinations of basal metabolic rate among various races are those of Benedict (57), Turner (565), Heinbecker (259), Steggerda and Benedict (538), and Williams and Benedict (613). The matter of individual differences in basal metabolic rate is considered in most of the general manuals on the subject. Knoll (344) has studied individual differences in the consumption of oxygen at work.

From the standpoint of working metabolism experimentation one of the most important physiological factors, subject to control in some instances and reportable in all cases, is the effect of differences in the extent of previous muscular training. Schneider, Clark, and Ring (505) have shown that such differences in the study of training have slight effect upon the basal metabolic rate, although there is a large number of studies indicating appreciable effect on work metabolism. Waller and DeDecker (591), studying carbon dioxide production, found indications of doubled efficiency in walking in the trained as compared with untrained state. Mague (552), in his contribution to the symposium on industrial psychology, found that persons accustomed to muscular work are able to eliminate higher concentrations of carbon dioxide, thereby economizing pulmonary ventilation. Briggs (113) has also noticed the effect of training in economizing pulmonary ven-

tilation through utilizing a higher percentage of oxygen from the air and excreting a higher percentage of carbon dioxide. He states that the normal percentage may be even doubled in the case of certain subjects. Simonson (522) offers evidence for the conclusion that exercise results both in an increased ability on the part of the muscles to remove lactic acid and a greater pulmonary economy in carbon dioxide elimination, and Liebenow (376) states that the effect of training is in the direction of increasing "the restitution constant." By this is meant that the oxygen debt is eliminated more quickly and the body reaches its normal carbon dioxide balance more rapidly. Hartwell and Tweedy (253) found only slight differences in pulmonary economy as between athletic and non-athletic women. Krogh (351) and Rosenheim (484) both speak of increases in mechanical efficiency as a result of practice in the task.

Attempts to study the physiology of this increase in efficiency have been made by Simonson and Reisser (523), an abstract of whose study is quoted from *Psychological Abstracts,* Volume I, No. 1315: "Training is physiologically defined as an increased output ability obtained by exercise. Recuperation is improved by repeated exercise of the function, although there is no reduction in the energy consumption with practice. This supports Reisser's view of training effects as due to an habituation to the toxin." Hietanen, Nikkinen, Nygssölä, and Sternberg (276) found an increase in efficiency during the first hour of walking on a slippery surface. This can probably be accounted for in terms

of the acquisition of skill rather than requiring a physiological explanation. Briggs (112) reports that "physical work is found by experience to be easier for unfit men when oxygenated air is breathed but no such difference is to be observed with fit men." This was not confirmed in the earlier study by Hartwell and Tweedy, already quoted, but would be theoretically significant if clearly established. Other studies of interest in connection with the physiology of training are those of Loewy and Knoll (384), Kaup and Grosse (316), and Knoll (344).

The experimenter using respiratory exchange determination methods in industry will ordinarily need to have but little concern for the effect of pathological conditions, but it is necessary to mention here that there are certain endocrine disturbances, such as hyperthyroidism, which may have a serious effect upon the total metabolism and that certain other disease states and atypical conditions should be observed and recorded if present. Bowen and Carmer (110) found that the added energy consumption for the obese as compared with the normal is only about what would be expected for moving the excess weight, although Wang, Strouse, and Smith (607) state that the heat production was greater in the obese than in normal or thin subjects and lowest in the normal subjects. They also found the mechanical efficiency greatest in the normal and least in the obese. Benedict (62) has studied individuals of unusual physical configuration, and Frank and Herzger (197) have studied oxygen consumption and res-

piratory quotient under various conditions of nutrition and bodily health.

Möller (428) has written an extensive study of basal metabolism in diseases of the thyroid gland and furnishes a bibliography of general nature, consisting of about 150 references, by author and journal. Boothby and Sandiford (102) report investigations indicating that thyroid cases require considerably more energy for a given piece of work than the normal. Others who have studied the effect of thyroid disturbance on work metabolism are Smith (526) and Curtis (151). The reference to Biedle (84) was retained in this bibliography through oversight, as it is of clinical significance only.

Peabody and Sturgis (460, 461) contribute hospital studies of heart disease of some significance to the student of respiratory metabolism. Langworthy and Barott (365) made the interesting discovery that, for five weeks after recovery from an attack of influenza of three weeks' duration, the energy expenditure per kilogram body weight was reduced by 4 per cent from the former requirement when doing the same amount of external work. McCann (413) found that the total pulmonary ventilation of five advanced tuberculosis cases was double that of the normal controls.

Owen, Cope, and Hill (454) called attention to a case of unsatisfactory metabolism test resulting from leakage of air into ears and Eustachian tubes as a result of perforation of ear drums. Such instances are not likely to occur in industry frequently enough to require special precautions. There are interesting and

sometimes surprising effects from other pathological conditions of the organism, but these will not ordinarily interest readers of this monograph. Those wishing further material concerning the effect of diabetes, various febrile states, epileptic contractures and various seizures, hynoptic rigidity, and other miscellaneous special conditions may be referred to the general manual by Grafe (225).

An important question both to the worker in basal metabolism and metabolism of muscular exercise is that of possible daily fluctuations due to undiscoverable organic factors. Such variations, if of great magnitude, would seriously affect the results of all types of metabolism experiments, since the factors producing them would be subject neither to observation nor to control. Wishart (621) has written on "The Variability of Basal Metabolism," and this has been abstracted in *Physiological Abstracts,* Volume 12, No. 1012: "The day to day variability in both metabolism and respiratory quotient (estimated by analysis of expired air collected in a Douglas bag) may be expressed by a coefficient of variation of four or five, i.e. the minimum and maximum of a series of observations in a single individual may differ by as much as 30 per cent. The variability is increased if the protein in the diet varies considerably, and decreased if the subject adheres to a very strict daily routine. Perfectly normal people may show basal metabolic rates of 20 per cent below DuBois standard." Hafkesbring and Collett (239) found daily variations of about 5 per cent in both directions with maximum variations about double that. Kunde

(358) also found indications of seasonal and daily variations. Lusk and DuBois (393) find that a man thirty or forty years of age may maintain a basal metabolic rate during eleven years within a variation of plus or minus 7.6 per cent. A sedentary existence was found to have the effect of reducing the metabolic rate. Harris and Benedict (252) have made a careful statistical study of the problem of variation in basal metabolism. It is interesting here to note that Hendry, Carpenter, and Emmes (268) found that post-absorptive oxygen consumption is uniform between 8:30 A.M. and 12:30 P.M.

Hafkesbring and Collett, quoted above, found basal metabolism 5 per cent higher in cold weather than in hot, and Griffith, Pucher, Brownell, Carmer, and Klein (233) verified this at least in part in finding oxygen consumption and pulse rate higher in winter than in summer, although carbon dioxide production showed no seasonal change. On the other hand, Gustafson and Benedict (237) report that "the average values for the oxygen consumption strongly suggest that the metabolism tends to be at a low level in the winter and to rise to a higher level during the spring and summer." Rowe and Eakin (491) have accumulated evidence that there may be a metabolic curve both in men and in women associated with some sort of sexual gland cycle. Lindhard (381) links seasonal periodicity in respiratory exchange to the varying intensity of the sunlight.

Another organic factor which might, at first thought, be supposed to play an important part in total meta-

bolism is that of menstruation. It appears that the influence of menstruation upon metabolism has been the subject of controversy and experimental disagreement for many years. The most definite conclusion one can reach from studying the literature is that, whatever differences there may be in metabolic rate brought about through menstruation, such differences are within plus or minus 5 per cent and may probably be considered negligible in ordinary industrial experiments. Kunde (358) and Hafkesbring and Collett (239) found a slight lowering in basal metabolic rate during the first few days of menstruation, and Rowe and Eakin (491) found a metabolic rise in the premenstrual week, while Blunt and Dye (87), in an extensive study, found no elevation in basal metabolism either before or during menstruation. King (321) quotes several late studies and regards the question of the influence of menstruation on metabolic rate as still being open in 1924. DuBois (172) also finds the question still open in 1927. Benedict and Finn (76) review a literature of about twenty titles in 1928, however, and report indications in their own experiments of a slight lowering in metabolism during the menstrual period. This investigation included one study of twenty subjects and two studies of one subject each.

The investigator in industry is, of course, interested in these studies of the influence of menstruation on basal metabolic rate only in their indirect bearing on the question of whether physical work proceeds at an undue energy cost during this period. Wiltshire (618) measured oxygen consumption and carbon dioxide out-

put during light work and found that the cost of work and rate of recovery were the same during menstrual and inter-menstrual periods. Bedale (43), in a recent intensive study of one subject, found differences in the energy cost of work during and between menstrual periods, although these differences were slight. The differences, surprisingly enough, were in the direction of increased efficiency during menstruation. "There seems no reason to think that the fundamental physiological rhythm of women is such as to affect, either considerably or constantly, the quantity or quality of their industrial work provided always that no pathological conditions are present." This study is supplemented by a bibliography of fifty references, with a review of the work done on the problem to date.

Physiological Factors Subject to Experimental Control. There remains now for discussion a large group of organic factors which are usually subject not merely to observation by the experimenter but are also more or less under his control. The necessity for controlling certain of these factors in metabolic experiments is universally recognized, although certain others are of minor importance and have not received wide attention. Most of the experimental work which has been done has been performed with a view toward determining the effect these factors have upon basal metabolic rate rather than possible influences upon the energy cost of muscular work. The experimenter who is interested in the latter must therefore rely partly upon indirect evidence from experiments of the former type.

Wishart (619), for example, has studied the influ-

ence of previous muscular activity and other factors upon the basal metabolism, but this does not tell us what effect previous muscular activity has upon the energy cost of work. His study is described in *Physiological Abstracts,* Volume 12, No. 1012: "The effect of an hour's moderately severe work on the previous day is to raise the basal metabolic rate one to two per cent and to lower slightly the respiratory quotient; these differences are so small as to be entirely obscured in only occasional observations."

The influence of sleep on basal metabolism in children has been studied by Wang and Kern (606), who found a drop in heat production ranging from 5.7 per cent to 30.6 per cent. Other studies of the effect of various degrees of muscular tension and relaxation will be discussed in a later section.

The effect of loss of sleep has been studied by Laird and Wheeler (361) and Landis (362). Laird and Wheeler, using the Douglas Bag method, reported an increase in the energy cost of mental work following loss of sleep but found no effect on errors and even found an increase in rate. Three subjects were given practice in mental multiplication for several weeks, or until they had reached their apparent limit of practice. Observations were then made for one week during which subjects were given eight hours sleep, and this was followed by observation for one week in which the subjects slept only six hours. There was, unfortunately, no return to the eight-hour sleeping schedule nor were adequate controls run over the period to permit complete isolation of the cost of mental work fac-

tor. As will be seen in the section on *Mental Work*, it is dubious whether reliable conclusions may be drawn from an experiment in which the cost of mental work is estimated by means of respiratory exchange in such short experimental periods; particularly is this so when the sampling occurs only after the close of the mental work.

The influence of various states of under-nutrition and fasting has been repeatedly investigated and the literature has been thoroughly covered by Morgulis (433), whose monograph carries a bibliography of more than one thousand references with complete titles. There are good author and subject indexes and the volume is written with strict adherence to the experimental approach. There are liberal quotations of data and charts of experimental results. Grafe (225) also supplies 78 titles on hunger as well as extensive bibliographies on other controllable factors. A classic experiment in this field is that of Benedict, Miles, Roth, and Smith (78). This and other related Nutrition Laboratory studies are discussed by Benedict (53) in a lecture on the social and economic implications of the ability to resist fasting. Kunde (358) found a temporary increase in the basal metabolic rate in prolonged fasting, and Landis (362) reports upon the metabolic effects of fasting in his study of the emotions.

The disturbances of metabolic rate caused by the taking of food has not always been eliminated from experimental investigations. Certain early experiments, such as that of Carpenter and Benedict (126), were greatly weakened in the significance of their con-

clusions through lack of having controlled the effect of food. Benedict and Murchhauser (79) reported in 1915 that "the heat production per unit of work is practically independent of the taking of food." This view has been seriously modified in recent years, so that Cathcart and Orr (136), for example, report that, although they had considered the food-taking element in their experiment as being rigidly controlled, in searching for the cause of what appeared to be errors in technique they discovered that the discrepancies were due to the practice of some of their subjects of partaking of midnight lunches.

Higgins (277) in 1913 demonstrated an increased alveolar carbon dioxide tension throughout the duration of the digestive processes and many studies before and since have corroborated the fact that digestion in itself does increase the metabolic rate. The effect of digestion on metabolic rate, whether the body is at rest or at work, depends in large measure upon the nature of foodstuffs concerned. Orr and Kinloch (452) summarize the earlier experiments and report an experiment of their own employing the Douglas-Haldane method. Their conclusions were: (*a*) Following a high protein meal the increase due to work is greater than in the preceding post-absorptive state. (*b*) Following a high carbohydrate meal the increase is less than in the preceding post-absorptive state. (*c*) Following a high fat meal there appears to be a summation of extra energy due to the food and that due to work. The ingestion of carbohydrate has been studied in the resting condition by Benedict, Emmes, and Riche (74)

and during work and rest by Cassinis (129). Viale (574, 576) has also studied energy consumption in human work before and after meals and during fasting. Wishart (619, 621) demonstrates in each of the references quoted that a high protein diet markedly increases the basal metabolic rate on the subsequent day, and Cathcart and Burnett (133) show that "under the conditions of the experiments the differences in the oxygen demand during work on diets which contain meat and those which are meat free are definitely significant" (low work respiratory quotients in the latter). General treatments of the specific dynamic action of food will be found in the references by Jahn (304) and Lusk (387).

Experimental researches toward finding a breakfast which will not interfere with the determination of basal metabolic rates should also be of significance in certain types of work experiments. Soderstrom, Barr, and DuBois (529) fed 30 grams of bread, 8 grams of butter, and one cup of caffeine-free coffee with 10 grams of sugar and 60 cc. of milk to five normal subjects preceding the determination of their basal metabolic rate. A slight rise was produced in respiratory metabolism, which disappeared after the third hour. Benedict and Benedict (45) also describe a nearly non-protein meal which is light and non-stimulating and suitable for feeding to subjects before a metabolism experiment. Bauer and Blunt (36) conclude that basal metabolic rate determinations of children may be made at noon, provided only that the breakfast does not exceed 420

calories, including 14 grams of protein, and that it is eaten at least four hours before the test.

The work that has been done on the effect of emotion is still too meager to show consistent and conclusive results; we know that any emotion may have an influence upon metabolism, but we do not always know in what direction to expect the change, nor how confidently to look for it. Thus Landis (362), experimenting with human subjects to discover the effects of fasting, insomnia, electrical stimulation, anger, etc., on metabolism, concluded that emotional disturbance, *per se*, does not lead to changes in the metabolic rate which are always in the same direction or of the same magnitude. Conversely, changes in metabolic rate cannot be considered as direct measures of emotional disturbance or cumulative emotional upset. On the other hand, Zeigler and Levine (622), prompted by the presence of abnormally high metabolic rates in certain psychoneurotic individuals, experimented upon the effect of emotion in such patients. Their results showed that these psychoneurotics usually respond when thinking about an emotion-producing aspect of their past history by an increased metabolic rate. Segal, Binswanger, and Strouse (511), in studying the nervous symptoms in exophthalmic goiter, found that individuals with toxic goiter not treated with iodine show the possibility of a dangerous emotional metabolism rise. Totten (563) determined the oxygen consumption before, during, and after a wide variety of intense emotional stimuli. He found no increase in half the cases studied and some increase in the others. Hafkesbring

and Collett (239) found that harsh or sudden noise produced a rise in metabolism of 10 per cent for between 5 and 20 minutes.

The possible effect of minor emotional disturbances has not always been realized. Early studies on mental and muscular work, even under such careful workers as Benedict and Carpenter (69, 126), were often inconclusive because of failure to take into consideration such potentially emotional stimuli as the influence of novelty. It may be that this is not a highly serious matter, however, for Hendry, Carpenter, and Emmes (268) have shown that unpracticed subjects differ but slightly from practiced, although one practice period is recommended.

McDowall and Wells (420) hold that monotony in the physiological sense is the exact opposite of emotion. They have established that vascular reactions cease as soon as the stimulations or effort become monotonous.

Grafe and Mayer (227) and Grafe and Traumann (229) have conducted experiments on the metabolic effect of emotions in the hypnotic state.

DuBois (172) has a good chapter on the effect of emotional states, and Grafe (225) has an extensive bibliography on emotional states and mental disorders.

Poffenberger (467) points out that metabolic techniques open a promising means of investigating the actual effect of drugs upon human efficiency. Without the use of such measures we can only determine that a certain drug has a certain effect upon an individual's performance in mental tests or other objective tasks.

Very often this effect is found to be either insignificant or else highly variable. If the energy cost of the task could also be measured before and after administration of the drug we would have a means of determining more accurately what actual influence the drug may have had upon efficiency. Not much work has been done along the line of this suggestion, however, so that we shall need to be content with mentioning a few of the more general reviews of the effect of drugs on basal metabolic rate.

Boothby and Rowntree (101) conclude that the common drugs (not including preparations representing internal secretion) do not demonstrably influence the basal metabolic rate, but that iodine, adrenalin, etc., do have a calorigenic action. Hardikar (250), however, found a tendency toward increased ventilation, respiratory exchange, and heat production with doses of quinine up to two grams, and Higgins (277) found a fall in alveolar carbon dioxide tension after administering coffee without food. A very extensive review of the effects of various drugs and poisons upon metabolism has been written by Barbour (32), and chapters on the effect of drugs will be found in the books by DuBois (172), McCann (412), and Amar (10).

The delicacy of respiratory exchange techniques is well illustrated by the fact that they register changes resulting from slight differences in body posture and various minor muscular movements. Experiments fairly consistently show an increased metabolic rate in standing or sitting positions as compared with reclining or relaxed positions, as well as increases from cer-

tain minor body movements. These results, while of little direct significance to the worker in muscular metabolism, are of importance in that body posture and minor movements may cause serious errors in metabolic experiments designed to study the effects of emotion, mental work, or other factors having an effect of low magnitude.

Benedict and Benedict (64) found that oxygen consumption increased 0 to 11 per cent in sitting as compared with lying, averaging about 3 per cent, and 9 to 24 per cent in standing as compared with lying, averaging about 10 per cent. They also found that one movement of the hand to the forehead per minute while lying was of no significance but that there was an increase of oxygen consumption of 11 cc. for each time the legs were crossed while lying down. For a more complete treatment of this material see Benedict and Benedict (67).

More energy is consumed, apparently, in lying flat in bed than when in a semi-reclining position. Emmes and Riche (185), experimenting with two subjects, showed that there is an increase in metabolism of 8 per cent in the lying position as compared with sitting upright, and Higgins (277) found a higher alveolar carbon dioxide tension in subjects in a relaxed position than when they were in an erect position. Soderstrom, Meyer, and DuBois (530) studied four normal men and two cardiac patients lying flat in bed and in the semi-reclining position propped up with a back rest, or else in a steamer chair. Twenty-one experiments

showed that metabolism averaged 3 per cent lower in semi-reclining posture.

DuBois-Reymond and Peltret (99*a*) concluded that energy expenditure is affected by the state of tension, and Wishart (619), as we have quoted above, even found that an hour's moderately severe work on the previous day resulted in a 1 to 2 per cent rise in the basal metabolic rate. Other studies have been made by Simonson (520) on the physiology of standing, by Liljestrand and Wollin (379) on body posture, and by Turner (564) on reclining, sitting, and standing positions. Benedict (79) and Studer (548) have measured the effects on metabolism of resting and standing as a preliminary part of their studies on muscular work. Similar preliminary studies of the resting state have been made by a large number of investigators.

Before concluding our discussion of the controllable organic factors we may mention a few miscellaneous influences which have not attracted such wide experimental attention. Glandular activity, for example, must take place at a metabolic cost of some sort, although we do not know in many instances just what such costs amount to. Bircher (85) performed an experiment the results of which implied that sweating does not in itself increase oxygen consumption to any great extent. The following is quoted from the *Journal of the American Medical Association,* Volume 80, page 729. "Bircher produced sweating in 20 persons by heat and light. The oxygen consumption increased 13 per cent over the basal rate and the temperature of the body about one degree centigrade. The increased

metabolism lasted longer than the sweating." Kost (347) found that oxygen consumption is not influenced through massage. Benczúr and Berger (44) found a slight increase in alveolar carbon dioxide tension with the application of heat and a slight lowering with the application of cold. This was a clinical study referring to local applications of these temperature differences. Pemberton (462), in a study of exposure of the body to the therapeutic application of external heat, found heightened blood flow, increased metabolism, and increased elimination of acids, chiefly through expiration of carbon dioxide. Lusk and DuBois (393) found that "cage life" reduces basal metabolism in the dog and has a probable counter-part in lack of exercise and indoor confinement in man. Ponzo (470, 471) studied the influence of volitional factors on respiration. He discusses psychological influences upon the rate and character of breathing, finding momentary suspension of breathing during close attention, acceleration and retardation from thoughts with varying affective tones, and influences from the slight laryngeal movements involved in thinking. These studies, of course, refer to the effect upon the respiratory curve and not upon metabolism as measured by respiratory exchange.

Mental Work. The subject of mental work will be given consideration here under its own specific heading because it is one of the most interesting and significant of the organic factors under control of the experimenter. It is placed last in the list, however, because its actual influence upon metabolic rate is so slight that it can be detected only in the most carefully controlled

experiments, utilizing the most sensitive techniques. It appears to be well established that, while purely intellectual work may have a measurable physiological cost, such cost is of such low magnitude as to make it easily obscured by practically any or all of the physiological variables listed in the preceding section. This makes the problem of providing adequate controls a most difficult one and excludes all but the most refined techniques of measurement. Routine determinations of the energy cost of mental work in business and industry will never be possible by means of respiratory exchange determination methods.

The classic experiment on the determination of the energy cost of mental work is that of Benedict and Carpenter (69). These authors reviewed the early work on metabolism in mental and muscular work and reported their own studies on this subject in detail. Although their experiment is usually quoted as demonstrating absence of measurable energy cost in mental work, their own conclusions are actually not quite so negative. "From the results. . . . it would appear that the pulse rate was slightly increased, the body temperature somewhat higher, the water vapor output increased by about 5 per cent, the carbon dioxide production increased by about 2 per cent, the oxygen consumption increased about 6 per cent and the heat production increased by about one-half of one per cent as a result of sustained mental effort such as obtains during a college examination." The reason this experiment is usually quoted as demonstrating no increase in metabolism during mental work is found in the fact

that these authors qualify each finding by pointing out possible inaccuracies of technique and limitations in interpretation. They disclose, for example, the serious oversight that the controls were all run after the mental work so that the element of novelty in the situation was absent in the controls, but present in the experimental trials. This is the principal factor tending to invalidate the conclusions of their experiment. A popular presentation of their results was written a year later by Benedict (48).

Other metabolic studies, showing the contradictory results characteristic of mental work investigations, are those of Ilzhöfer (300) and Chlopin, Jakowenko, and Wolschinsky (139). The former, using the Krogh apparatus, found that metabolism during mental work is not increased materially, while the latter study indicates an increase in metabolism in general. A review of Ilzhöfer's work will be found in the *Journal of the American Medical Association* (83). Chlopin, Jakowenko, and Wolschinsky review and criticize the earlier literature on mental work in a way that gives support to their own findings.

Becker and Olsen (39) have described their methods and apparatus in great detail in the expectation that other psychologists might wish to use them in similar experiments. Poyer (473) has written a theoretical article with a brief bibliography on mental work. Spencer (533*a*) contributes summaries and references on mental work, and Grafe (225) has a bibliography of 23 titles on this and related subjects. The DuBois manual (172) has a chapter on mental work, and fur-

ther reference material will be found in Takahira and Ishibashi (554), Mainzer (399), Liebermann (377), and Kestner and Knipping (318). This latter study is on the relation of protein diet to mental work.

The influence of mental activity on vasomotor processes has been studied by several workers. Gillespie (215), in a study of the relative influence of mental and muscular work on the pulse rate and blood pressure, concluded that mental work produces an increase in pulse rate and blood pressure, that this increase is independent of emotional factors, and that it cannot be accounted for by movements of articulatory muscles or known muscle tensions. Combined mental and muscular work was found to produce a greater effect than either singly. In women the pulse rate change was found to be proportionately twice as great as the blood pressure, but with men it was about the same. McDowall (419) also found that pulse rate and blood pressure are both raised by either mental or muscular work. This study carries a good non-technical account of the physiology concerned. Day (157) gives an excellent summary of experiments on effect of mental work on pulse rate and blood pressure, as well as other factors influencing vaso-motor processes.

Dodge (165) gives a thoughtful account of the psychological justification for including a study of mental work in psychology. In this article will be found a critical discussion of pulse rate determination as a simple metabolic rate determining technique, and a statement of the desirability of finding and developing a more reliable index which will retain the advantage

of simplicity. Dodge describes elaborate electrocardiograph and string galvanometer experiments in connection with the study of the metabolism of mental work.

Several other techniques have been employed to determine the energy cost of mental work. Goldberg and Lepskaia (219) have studied the alterations of the white corpuscles in the blood during mental and physical work. They found more pronounced changes during mental than physical work, with a return to normal in about two hours. Knipping (335) has found the measurement of phosphoric acid in the blood a more satisfactory measure of mental work than respiratory exchange determination.

Environmental Factors Affecting Metabolism. On casual thought it would appear to be a simple matter to classify the factors influencing metabolism as being either environmental or organic. The effect of food and drugs is obviously physiological; the influence of temperature extremes in the atmosphere is clearly environmental. But when we attempt to classify such factors as oxygen lack in the atmosphere we find ourselves in some degree of doubt, and when we go to such factors as the effect of noxious gases, as studied by Henderson and Haggard (266), we realize the relativity of our classifications. We classify various factors as an aid in discussing them, but these classifications do not carry the implication that there is a fundamental distinction between the groups comprising our artificial schema.

The effect of high and low atmospheric temperatures

is probably the most important of the environmental variables affecting metabolism. This is treated, of course, in the general manuals of which Lusk (387) may be taken as an example. Lusk publishes on page 149 an interesting table showing the influence of clothing on metabolism. Carbon dioxide production is increased when scant clothing is worn in a cold room. McConnell and Yagloglou (416) and McConnell, Yagloglou, and Fulton (417) demonstrated that the rates at which oxygen is consumed and carbon dioxide produced increases with exposure to either high or low temperatures, and that the metabolic rate increases rapidly when the environmental temperature is higher than that of the body. Other studies on the effect of high air temperatures on basal metabolism were made by Benedict, Benedict, and DuBois (46), and Bircher (85). Moss (436), Volshinski, and Yakovenko (579) and Viale (572) have studied the effect of high temperatures on working metabolism. Hill (288) and Hill and Campbell (289) studied the effect of cool air currents on working metabolism, the latter finding that "during 15 minutes work on a bicycle ergometer the cooler conditions greatly relieved the heart—reduced the pulse rate—although the gross and net efficiencies of muscular contraction were not affected." This is quoted from the *Abstract of Literature of Industrial Hygiene,* taken in turn from *Physiological Abstracts,* Volume 7, pages 472 ff.

The effect of temperature differences is certainly great enough to warrant considerable care on the part of the investigator in controlling this factor in experi-

ments upon muscular work. There is still some question as to just how nearly constant the temperature must be maintained in determining basal metabolism. There is some evidence that basal metabolism determinations should be made with the subject in a warm bath of constant temperature. This is discussed by Benedict and Benedict (66) and Mayer and Wurmser (410).

The effect of tropical and Alpine climates on basal metabolism and working metabolism is the long-time aspect of the temperature differences discussed above. The physiological effects of tropical climate have been surveyed in a comprehensive way by Sundstroem (551). Ozorio de Almeida (455) and Hindmarsh (294) found a lowered basal metabolism in tropical climates. Hindmarsh attributed this to the ready relaxation favored by the warm environment. Others who have studied the effect of tropical climate on basal metabolism are Coro (146) and Montoro (429). Cohen (140), in a letter to the *Journal of the American Medical Association,* pointed out what he considered an old and important observation made by Mayer in Java indicative of lessened oxidation of tissues in tropical climates.

It is often difficult to separate the effects of race and the effects of climate in studying metabolic differences of peoples living either in tropic or frigid regions. Heinbecker (259), for example, finds that "the basal metabolism of Eskimos is considerably higher than that of persons living in temperate zones." But we cannot be certain about the factors to which this should be attributed. Hill and Campbell (290) found that the

heat production of resting subjects exposed to Alpine air increased between 40 and 50 per cent in the case of clothed adults and 60 to 90 per cent in the case of nude children. Viale (571) has studied working metabolism in an Alpine climate. Corlette (145) found a cold, moist atmosphere to be a milder environment than cold, dry air. Dhar (159), in a theoretical article concerning man only indirectly, notes that it is more advantageous for a man living in a colder climate to change to a warmer climate than to change in the contrary direction.

There are various types of radiation, such as sunlight, Roentgen rays, ultra-violet light, etc., which have been thought to play some part in determining metabolic rates. Their influence, however, is not great and this factor will probably not need careful control in industrial experiments. A review of the literature on sunlight and other kinds of radiation has been written by Laurens (368). There is a section on the effects upon metabolism, pages 40-50, and a bibliography on the effects upon metabolism, page 87. Ultra-violet radiation has been studied by Crofts (149) and by Mason and Mason (408). The latter found that ultra-violet light from a quartz mercury vapor lamp is capable of lowering the total metabolism of some persons. Flickinger (193) and Campbell (120), however, are unable to find this effect. Lindhard (381) believed that seasonal periodicity in respiratory functions is due to the varying intensity of the sunlight.

Lyon and Greisheimer (394) describe an experiment in which the air surrounding the body of human sub-

jects was kept constant in temperature and humidity while air of constant temperature and varying humidity was inspired for 10- to 15-minute periods. Their tentative results indicate higher pulse rate, lower respiration rate, higher arterial pressure, and greater peripheral vascularity while breathing moist air than while breathing dry air.

A good review of the work on high and low atmospheric pressures, as well as with air of abnormal composition, will be found in Haldane (243). The effect of low pressures has been studied both in specially constructed chambers and in experiments performed at high altitudes. A celebrated experiment of this type is that of Douglas, Haldane, Henderson, and Schneider (169). This experiment, known as the Anglo-American Pike's Peak Expedition, was productive of many useful conclusions. The respiratory exchange was found to be unaltered whether at rest or at work, and acclimatization was distinctly evident after two or three days' residence on the summit. Schneider (501) reported changes in the blood circulation and respiration of a man who lived for a long time on the summit of Pike's Peak. The influence of rapid change of altitude on circulation and respiration has been studied by Hecht (257). Other altitude studies are those of Kestner and Schadow (319) and Herxheimer, Wissing, and Wolff (273).

Schneider and Clark (502-504) have conducted a series of laboratory experiments on the effect of low barometric pressure on working metabolism. These are careful experiments in which several physiological fac-

tors are studied, although work costs are estimated only on the basis of oxygen consumption. They point out (504) that when one is first exposed to low pressure the oxygen want stimulates the respiratory center, causing a blowing off of preformed carbon dioxide. This cannot therefore be used for measuring metabolism in experiments under low atmospheric pressure. Schneider, Truesdell, and Clark (507) studied oxygen consumption at rest during short exposures to low pressures in an experiment preliminary to those just described.

There still remains a wide variety of environmental conditions which have been thought to play a part in determining metabolic rates, but most of these are of small significance to the industrial investigator. Some of them are principally of therapeutic interest, such as the study of massage by Kost (347) and studies of therapeutic baths, such as the carbonic acid bath studies by Laquer and Gottheil (367) and Wassermann (608). The carbonic acid bath increases the rate of carbon dioxide exhalation. The effect of sea baths has been studied by Margaria (401), who found an influence on the temperature of expired air and the ventilation of the lungs.

Benedict and Finn (75) publish results of experiments showing that the basal metabolic rate of an individual is in general sufficiently fixed as to be unaltered by a summer's vacation, even when pronounced subjective impressions of regeneration are experienced.

McDowall and Wells (420) have written a theoretical discussion of the physiology of monotony. This has

already been treated in our consideration of the effect of emotions.

Diserens (163) has made a comprehensive study of the influence of music on behavior, but, although the volume carries an exhaustive bibliography, there are only two references to the effect of music on gaseous metabolism and both of these are very early animal experiments. The reviewer has encountered no experiments up to 1928 on the effect of music upon human gaseous metabolism. This would appear to be a fruitful field for investigation both from the theoretical side and from the standpoint of a possible reduction in energy expenditure during work.

General References on Factors Affecting Metabolism. In concluding, we may mention several references in which the factors influencing metabolic rate are reviewed in a general way, including a consideration of both organic and environmental factors.

A general survey of factors affecting metabolism was written by Benedict (50) in 1915 and a very readable popular account in 1928 (63), in which the history of metabolism and its measurement were also included. He wrote a more detailed treatment of the factors affecting metabolism as well as a description of types of apparatus, etc., in 1924 (54). Benedict and Carpenter (70) discuss a variety of factors in a series of experiments on metabolism of healthy men in the resting state. DuBois (172) lists six variable factors which must be controlled in metabolism experiments and twenty factors to be taken into account in interpreting results. Knipping (334) has also discussed the many

factors which must be considered in basal metabolism investigations. Amar (10) treats of these factors more particularly as they deal with working metabolism. Other studies of a more special sort are those of Sainton and Péron (495) and Harris and Benedict (251).

II

APPARATUS AND METHODS

General Manuals on Apparatus and Methods. Aside from a few very early contributions such as those of Jaquet (306), Loewy (383), and Tigerstedt (561), all of the general reference manuals, monographs, and textbooks that treat of methods and apparatus used in determining human energy consumption have already been mentioned in preceding sections of the present review. It should be useful, however, to assemble them in a chronological order here as a matter of convenience.

Those interested in early types of apparatus will do well to investigate Johansson (309), Lefevre (373), and Amar (10) in addition to the three studies mentioned in the preceding paragraph.

A more modern group of references comprises those of Carpenter (123), published in 1915, Krogh (354), published in 1916, Boothby and Sandiford (104), published in 1920, and Sanborn (496), published in 1922. Some of these books will be referred to later in discussing specific types of apparatus.

Aside from the textbook by DuBois (172), already quoted several times, all of the modern reference manuals in which there is adequate treatment of methods and apparatus have been published in German. Benedict (54) has written an article in which all the forms of apparatus used by the Nutrition Laboratory are described. These include chamber types, portable

types, student apparatus, micro-respiration apparatus and devices such as ergometers and tread-mills. His primary concern, however, is with the apparatus used in determining basal metabolism. Klein and Steuber (325) treat fully of the various types of apparatus used in determining oxygen and carbon dioxide content of gases. Abderhalden (2) edits an exhaustive and comprehensive symposium on calorimetry and gaseous metabolism providing a complete working manual on all types of apparatus and methods. Krauss (348) describes and illustrates a wide variety of experimental apparatus and methods, although he does not include much that is useful in work experiments. Knipping and Rona (343) describe the best modern methods of both direct and indirect calorimetry, tread-mills, ergometers, masks, electrical devices and many modern improvements and useful points of technique. The sections on energy metabolism, pages 98-125, and work metabolism, pages 223-240, constitute a clear and comprehensive account of modern techniques, well organized, and supplied with splendid cuts and diagrams. Knipping and Kowitz (342) have written a splendid little working manual with illustrations of apparatus and descriptions of apparatus, both early and modern. Their interest is not primarily in working metabolism, however.

In addition to the various books listed above there are several articles in the periodical literature of considerable interest here. Benedict (58) has written an extended and useful account of modern techniques in the measurement of the gaseous metabolism of man,

and will be referred to later in discussing certain of the methods treated. Hendry, Carpenter, and Emmes (268) describe an important experiment in determining the relative merits of various combinations of respiration apparatus and breathing appliances. Pickworth (465) has also compared different techniques, finding that basal metabolic rates determined in a chamber were more than 20 per cent lower than those determined by the Douglas Bag method.

Other journal articles of less significance to the industrial worker are those of Meserve (425), describing certain types of metabolic apparatus, and Stoner (541), describing the organization of a metabolism laboratory suitable for a large hospital. The reference to Jones (312) is only of minor significance and the reference to Cowgill (147) was retained in this bibliography through oversight, as an animal experiment is described which has no interest for us.

Incidental and Non-Quantitative Indexes of Metabolic Rate. Before discussing apparatus and methods used in respiratory exchange determination as a means of discovering the energy cost of work, we shall do well to examine briefly a few of the more important physiological indicators of metabolic rate which may be employed in experiments that do not demand quantitative results, expressible in terms of calories per unit of time.

The most important and the most widely used of such measures are vaso-motor records such as pulse rate, pulse pressure, blood pressure, etc. Gillespie (215) experimented upon the relative influence of mental and

muscular work on the pulse rate and blood pressure, having been led into this field through Benedict's assertion that "pulse rate indicates, in a general way, internal muscular work and muscular tonus." He found pulse rate and blood pressure to be affected by mental work or muscular work or both. Cathcart, Bedale, Macleod, Wetherhead, and Overton (132) found a general, but not strictly proportionate, correspondence between pulse rate and gaseous exchange and blood pressure and gaseous exchange. McConnell and Yaglogløu (416, 417) also found that both pulse rate and body temperature correlate fairly well with the basal metabolic rate, and Read (479), as a result of 300 determinations, concluded that pulse pressure and rate vary in the same direction as the basal metabolic rate. Jackson (303) even contended that the two factors, pulse rate and pulse pressure, taken under basal conditions enable one to estimate the basal metabolic rate with considerable accuracy. Smith (525) found pulse rate a fair metabolic index in grade walking at various rates and upon different grades, although it was not reliable in the case of horizontal walking and even in grade walking showed wide individual fluctuations.

Hafkesbring and Collett (239), on the other hand, found no correlation between temperature and pulse nor between pulse rate and basal metabolic rate. Frumerie (199), furthermore, found pulse rate unsatisfactory as an index of muscular activity, and Dodge (165) considers pulse rate determination only a makeshift in mental work experiments. Pulse rate determinations are rather easily obtained in most experi-

ments, however, and record of them will be found in most experiments on working metabolism. A good summary of the experimental work which has been done on pulse rate and blood pressure under varying conditions has been written by Day (157).

A closely related physiological measure is that of deep body temperature. This was recorded in many of the experiments performed by Benedict and his colleagues, for example, the study by Benedict and Cathcart (72). There is usually a fair correspondence between changes in deep body temperature and the amount of external work being performed.

Knoll (344) has considered the relationship between respiration rate and the volume of air inspired, but there does not seem to have been much work done on correlating respiratory rate with external work. Schneider (499) finds that breath-holding power with the lungs deflated, when the subject is under basal rate conditions, yields results which harmonize closely with those obtained by means of the usual methods of determining basal metabolic rate. This is, of course, a wholly different matter from that of determining energy expenditure from respiratory rate, but the contribution may prove significant in helping to establish respiratory rate as an approximate index of energy metabolism.

The electrocardiogram has been used by Messerle (426) in experiments upon work and the effect of rest pauses. He reviews the literature in this field and indicates that such measures may provide a useful method of studying muscular work. Dodge (165) and others

much more recently have used the electrocardiogram for studies of mental work and emotions.

An approximate index which would appear to be very useful, but which has not been widely employed, is that of determining loss of weight during work. Shepard (515) and Moog and Schwieder (430) utilized this measure, Shepard using it in an industrial study and Moog and Schwieder in studying swimming and running. The latter investigators found that swimming 1000 meters caused as great a loss of water from the lungs and skin as was produced by running 10,000 meters.

A determination which is fairly simple to make but which is not highly significant is that of the acidity of the urine before and after work. Hastings (254) found that the urine of men engaged in manual labor tended to be of a slightly higher degree of acidity than that of men at rest, and that, while the urine of muscular individuals was more acid after work than before, there were variations in the case of weaker persons.

Many of the most complete and careful studies of muscular work have included records of most of the measures we have mentioned as well as determining respiratory exchange or its equivalent. Typical thorough studies of this type are those of Cook and Pembrey (144), McKeith, Pembrey, *et al* (395), Benedict and Carpenter (70), and Bedale (41). Although Bedale determined ventilation rate, respiration rate, pulse rate, and blood pressure, in addition to oxygen consumption, the pulse rate and ventilation rate only

were found to bear a significant relationship to oxygen consumption.

Direct Calorimetry. The interest of the industrial engineer or the industrial psychologist in the methods of direct calorimetry is principally an historical one. The many careful experiments of Benedict and Carpenter, for example Numbers 69, 70, 71, and 124 of our bibliography, as well as numerous experiments by other early workers have demonstrated beyond all doubt that energy expenditure may be as accurately determined indirectly through the respiratory exchange as by the most sensitive direct measurements in a respiration calorimeter. Those who may have an interest in this type of apparatus are referred to the studies just enumerated and to the somewhat more recent studies of Langworthy and Milner (366) and Langworthy and Barott (365). The differential calorimeter described by Noyons (450) is one of the few calorimeters in use at the present time. Noyons' complex and expensive direct calorimeter is described as being responsive to a muscular act on the part of a subject placed within it in about 10 seconds and sufficiently sensitive to measure the heat resulting from raising the forearm only 15 centimeters. The book describing this apparatus was unavailable to the reviewer; it is abstracted in *Psychological Abstracts,* Volume 1, No. 1497.

Some of the older types of calorimeters provided for determination of energy expenditure, not only by means of direct calorimetry, but also through some form of respiratory exchange determination, usually by means of closed circuit methods. Another characteristic of

the older types of apparatus was that many of them were constructed to accommodate studies of several persons participating in an experiment within the same chamber. The use of the closed circuit methods in indirect calorimetry will be treated in detail in a later section; we shall pause here for a few references concerning group calorimetry. Benedict, Miles, Roth, and Smith (78) describe a well-constructed group respiration chamber with tread-mill and facilities for work experiments, and Benedict and Johnson (77) describe a large chamber for studying groups of thirty to forty persons by carbon dioxide elimination only. Gullichsen and Soisalon-Soininen (236) and other Swedish experimenters have also employed respiration chambers with provisions for determining carbon dioxide output only. Benedict (58) describes a new respiration chamber of the open circuit type which should prove useful in careful laboratory studies of working metabolism. The sensitive gas analysis technique developed by Carpenter (122) is employed in conjunction with the chamber described by Benedict.

Indirect Calorimetry: Open Circuit Methods, Gasometer Type. The remaining sections of this part of our paper will be concerned with apparatus and methods used in indirect calorimetry; these will be considered under the two broad categories of open circuit methods and closed circuit methods, which will in turn be subdivided into various types.

Open circuit methods in their primary form may be thought of as consisting of a calibrated gasometer for the collection of expired air and suitable arrangements

for sampling this air for later chemical analysis. Some variant of this standard form is typically regarded as essential in scientific work in which no sacrifices are to be made in completeness or accuracy.

A good description, with student laboratory instructions, of the apparatus used in the gasometer or spirometer method has been written by Henderson (264). Haldane (242) describes the original Haldane gas analysis apparatus, now often modified after Henderson (263). DuBois (172) says of this method, page 93, "The technic of air analysis is rather difficult, requiring at least a solid month of practice." The present writer, however, following the instructions given by Bailey (28) and Boothby and Sandiford (104), was able to run analyses which checked satisfactorily within one week. It would seem that a person trained in laboratory methods should not require the amount of time mentioned by DuBois, although it is true that even a mechanically inclined experimenter is likely to have difficulty with the apparatus occasionally for a rather long period of time. The manual by Boothby and Sandiford carries full instructions covering all parts of the method necessary for ordinary experimental work. They describe the collection of gas samples, the analysis of gas, methods of computation, and use of accessory apparatus such as the gas mask. Bailey (28) has also written a very careful and detailed description of basal metabolic rate determination by the gasometer method. He describes the use of sampling bottles and illustrates each step in gas analysis.

A useful note with respect to the sampling bottles

this method involves is that by Carpenter and Fox (127). "The results show no evidence of stratification . . . . and the diffusion was found to be rapid and complete even under extreme conditions. . . . Once a homogeneous sample of air is drawn into the container, it may be analyzed subsequently at any time without mechanical shaking and without fear of inadequate mixing."

A description of these methods of analysis will also be found in Hawk and Bergheim (256).

In addition to the Henderson modification mentioned above, improvements of the Haldane apparatus have been made by Newcomer (448), Gmeiner (217), and Martini and Pierach (407). The two latter studies describe a mechanical mixing device to be used in connection with the Haldane apparatus. Stoner (544) describes and illustrates a blank for recording the data and calculations, without logarithms, of the determinations involved in the gasometer method. He also supplies formulae (543) simplifying the calculations.

The same open circuit principle with subsequent gas analysis is used in the recent respiration chamber described by Benedict (quoted in preceding section), and is also used in the cot calorimeter described by Grafe, Strieck, and Otto-Martiensen (228). The refined and delicate gas analysis technique employed was developed by Carpenter and described by him in 1924 (122). Grafe, Strieck, and Otto-Martiensen also describe and illustrate the Carpenter apparatus. According to Benedict, the Nutrition Laboratory has now definitely abandoned closed circuit chamber methods in its favor.

Krogh and Lindhard (357) also sponsor an open circuit method in conjunction with a closed room of known volume with accurate subsequent determinations of the small changes in the gas concentrations by means of a highly sensitive gas analysis technique.

Other developments from the Haldane apparatus, or, more properly, back toward the original Orsat device, must also be mentioned. Boigey (93) objects to Laulanié's apparatus on the grounds of nonportability and high price, to Haldane's on account of complexity, and to Waller's on account of limitation to carbon dioxide and the absence of correction or compensation for the effect of temperature. He describes the construction and use of his own highly simplified device which is an inexpensive apparatus mounted in a small carrying case, using potassium hydroxide and pyrogall. The device is manufactured and may be obtained from Pirard. McKinlay (421) simplifies gas analysis by the easy expedient of using only the sodium hydroxide side of the Haldane apparatus, thereby determining only the carbon dioxide percentage. Hindmarsh (294) compared basal metabolic rates determined on the basis of oxygen alone, carbon dioxide alone, and the two gases taken together. He found that using carbon dioxide alone tended to obscure certain important differences revealed in a more complete experiment.

A careful study and comparison of open and closed circuit methods was made some years ago by Hendry, Carpenter, and Emmes (268). They experimented with all possible combinations of the Benedict Portable

(closed circuit) and the so-called Tissot, or respiratory valve, open circuit apparatus, with mouth pieces, nose pieces, and half-face masks. Both types of apparatus and all three breathing appliances yielded a fair degree of accuracy for oxygen consumption but the mask gave the lowest carbon dioxide values and respiratory quotients of the three breathing appliances. Carbon dioxide output was found to fluctuate widely in the case of certain subjects when wearing any type of breathing appliance and was usually higher with the portable apparatus than with the respiratory valve type. A series of experiments on the reliability of duplicate trials indicated general physiological superiority of nose pieces and respiratory valve apparatus, but these were not recommended because of technical difficulties. The mask was favored for use with the respiratory valve apparatus and the mouth piece with the Benedict Portable apparatus. The respiratory valve apparatus with mask was found best when reliance was placed on carbon dioxide elimination alone.

Open Circuit Methods, Douglas Bag Type. Mechanically it is a simple matter to substitute a rubber bag for the gasometer or spirometer in the set-ups described in the preceding section. The suggestion that this be done was originally made by Douglas (167) in 1911 and the method has since grown to wide popularity under the name of the Douglas Bag Method. It has been widely used in industrial experiments such as that of Rosenheim (484) and in laboratory experiments of a more technical nature such as that of Cathcart and Stevenson (137). The original note written

by Douglas was sketchy and inadequate, but the method has been described in considerable detail by Cathcart (131). Cathcart supplies condensed instructions for the use of the Haldane analyzer and describes methods of computation, as well as details in the technique of using the Douglas bag.

One of the serious limitations of the Douglas bag technique as compared with the use of a gasometer is that the bag cannot be made of as large capacity as is possible with the metal bells. Krogh (354) reports that "with violent exercise a bag taking 60 liters will not hold the air expired during one minute." This limitation has been overcome by Campbell, Douglas, and Hobson (121), using a battery of Douglas bags, and Hill, Long, and Lupton (285), who describe a similar battery of bags for the study of rapidly changing metabolism. Furusawa (203) describes a spirometer modification making it possible to study rapidly changing gaseous metabolism over a long period of time as in the bag-battery method just mentiond. A fan is mounted to operate within the bell of a large galvanized iron spirometer and the volumes are read from a scale beside the falling counter-weight. This has the advantage of simplicity and the possiblity of making direct volume readings.

Raper (476) enumerates several disadvantages of the original valve supplied with the Douglas respiration apparatus and publishes a diagram of a new valve by Siebe, Gorman and Company in which these disadvantages are overcome. Kommerell (345) simplifies the calculations of the Douglas Bag Method through the

use of nomographic tables. Moss (436) used the Douglas Bag Method in a study of energy costs of work at high air temperatures but computed the energy expenditure, not on a theoretical basis as basal metabolic rates are determined, but by comparing results from the unknown task with results obtained on an ergometer study of the same individual. A physiological deficiency in his methods is seen in the fact that his subjects were required to inspire through a mouth-piece connected by a three-foot tube to a gas meter, and to expire through a three-foot tube connected with the Douglas bag.

Bedale (41) apparently used the Douglas bag in an experiment in which samples were collected for less than one minute and based her conclusions on determinations of the oxygen consumption alone. Her results appear to be highly significant, however, in spite of the fact that carbon dioxide production and respiratory quotients were not determined. A more serious curtailment in technique would have been the ignoring of oxygen consumption data and reliance upon carbon dioxide elimination alone. As will be seen from the section on the Waller Method there is use for such a simplification in certain industrial problems, but it should not be resorted to in investigations of a more refined sort.

There is still some question as to the reliability of the Douglas bag technique even when used in its complete form, so the experimenter should be very wary of simplified procedures in experiments in which the fluctuations of the variable being studied are of low

magnitude. We shall close our discussion of the Douglas bag method with quotations favorable to the method and others opposed.

Carpenter (123) finds the Douglas bag reasonably satisfactory for short periods of time provided the bag is of 100-liter capacity. He advises pneumatic nosepieces when the bag is used with untrained, disinterested subjects. Greenwood, Hodson, and Tebb (231) used the Douglas Bag–Haldane Method in studying the working metabolism of a fairly large number of women. They were able to verify the Douglas-Haldane technique in the sense of demonstrating uniformity of results among moderately trained analysts. Dautreband and Davies (155) are critical of closed circuit methods for clinical use and describe in detail the use of the Douglas-Haldane Method, which they prefer for use in the clinic. Hill and Lupton (287) employed the Douglas bag method in an experiment on muscular work in which they reported that a steady state was reached within two minutes.

On the other hand, Cathcart, Lothian, and Greenwood (134) discuss statistically the validity of Douglas-Haldane observations and publish a series of 96 observations of the same subject under uniform conditions. Twelve per cent of the cases were found to lie beyond three times the probable error. The "error" is compound, that is, it is dependent both upon experimental technique and physiological fluctuations. Pickworth (465) even found that basal metabolic rates determined in a chamber were more than 20 per cent

lower than those determined by the Bag method. He believed that this might be due to the lesser emotional disturbance and the increased possibility of muscular relaxation. Wishart (621) found that the maximum and minimum of a series of observations of a single individual by the Douglas Bag method may differ by as much as 30 per cent. It is also significant to note that Sargent (497) found that the total oxygen consumption during recovery cannot be estimated by observing partial recovery and applying a constant correction.

The Waller Simplification of the Douglas Bag Method. Having noted the limitations of the Douglas Bag method as well as its advantages, we are in a position to consider a simplification proposed by A. D. Waller several years ago and used by him in a number of studies of an industrial nature. The method was proposed only as an approximation method but was developed on the grounds that it provided the only possible way of studying respiratory exchange of industrial workers actually engaged in the performance of their duties. The method has been in disuse the past few years, and the present writer would hesitate to revive it were it not for the fact that the method, if its validity can be established even within low limits of accuracy, will provide the industrial psychologist and the industrial engineer with the most usable of all tools available for the study of working conditions, rest pauses, and related industrial factors such as were described in the introductory section of the present work. We shall enumerate the actual studies made by

Waller and then proceed to a discussion of the evidence for and against the validity of the Waller Method and the material available for judging the limitations to be imposed upon the interpretation of results secured from the method.

Because of the rather wide disregard Waller's methods have encountered there are many people in this country who are unacquainted with his work and do not associate the Waller Method with the name of the well-known physiologist who achieved greater distinction for some of his other contributions to science. In his obituary (604) we learn that he was the son of A. V. Waller, a celebrated English physiologist, and that in 1892 he was made a Fellow of the Royal Society, being at that time 36 years of age. He was the first to obtain a human electrocardiogram, accomplishing this in the eighties, long before the introduction of the string galvanometer. He was also well known for his later work with the psychogalvanic reaction, as well as for his sponsoring the method of energy consumption bearing his name. The method met criticism during his own later years, but death cut off his project to establish the validity of the method in a carefully controlled experiment of some magnitude.

Most of Waller's own reports on his experiments are brief in nature and there is more or less repetition in the description of method. In an early study (584) he states that he uses a 20-liter rubber pillow fitted with a valve and two-way tap, opened and closed at a signal from an assistant with stop-watch. A mouth-piece and nose-clip are used, the bag being held by the subject.

Expired air is collected for a period of only 30 to 60 seconds, the workmen's activities being interrupted for that period. In this paper Waller supplies statistics and a graph showing carbon dioxide ordinates at one-minute intervals before, during, and after stair climbing in justifying his procedure of collecting the expired air *immediately* upon interrupting the men's work. The energy consumption calculations are based upon carbon dioxide determinations, with an assumed respiratory quotient of about 84. Details of the computation are given in another article in the same volume of the same journal (582), and he also explains his method of assigning mechanical work equivalents to volumes of carbon dioxide. He contends, for example, that, at an R.Q. of 84, one cubic centimeter of carbon dioxide is equivalent to 2.5 kilogram-meters. This description of his method seems to have been favorably reviewed by Stiles in the *Abstract of the Literature of Industrial Hygiene,* Volume 1, page 41.

Data useful in preparing graphs of carbon dioxide output over short periods are supplied in a later contribution (587). In another paper (596) he proposes a chart defining various types of work in terms of carbon dioxide discharge. Sedentary work, for example, is described as work in which less than five cubic centimeters are expired per second, this signifying an energy consumption of less than 100 calories per hour. At the other end of the scale is placed the heaviest of manual labor, with a carbon dioxide exhalation of more than 20 cubic centimeters per second, or an energy consumption of over 400 calories per hour.

Articles 585 and 595 in our bibliography are replies to objections to his method and will be referred to again in the next section. In another study (598) he reports an experiment performed upon one subject in which energy consumption was estimated by the usual method and by the Waller Method. A reasonably accurate correlation between the two was found. The method is defended further (602) in his contention that the error of the Douglas-Haldane method of measuring physiologic cost is plus or minus one per cent and that under the method of measuring the carbon dioxide alone for brief periods, as he proposes, it is increased to only plus or minus 5 per cent, provided care is taken to make all measurements after a steady regime is established. By doing this there is practically no interruption of the industrial processes on which the worker is engaged and the sampling proceeds practically under actual conditions of work. Many samples per day may be taken by one operator and if the results are treated statistically it is possible to derive conclusions of significance in industry, since the effect of at least some of the variables of industrial interest is likely to be of a magnitude at least three times this probable error.

Readers of this and other articles will find that Waller's curves all show hourly increase in carbon dioxide during each period of work, a restorative effect in the dinner hour, a rapid fall in carbon dioxide discharge during a rest pause of three or four minutes, and a higher maximum in the afternoon than in the morning. In this same paper he discusses the physiology of this progressive increase in carbon dioxide

discharge in continuous muscular work and mentions his intention of treating the subject in greater detail in a later paper. This apparently was never done. Waller himself regarded the phenomenon as a progressive increase in the energy cost of continuous work, although other physiologists have denied the physiological effect and expressed doubt concerning his explanation.

Waller believed that his method could be used in estimating energy cost of all types of muscular work from the heaviest to the lightest. Numbers 581, 583, 589, 601, and 603 of our bibliography are all reports of studies of heavy work such as dock labor and work in cold-storage plants. In these articles will be found repetitions of his descriptions of method and computations. His contention that energy cost of work is increased at the close of a long spell of uninterrupted work was supported in his observation (589) that the cost of walking a given distance was doubled at the end of a working period as compared with the beginning of the same period.

Studies 586, 591, 597, and 600 of our bibliography are all concerned with the energy cost of walking and marching. There are further explanations in these references of his experimental and computation methods.

Studies 592, 593, and 594 are concerned with the physiological cost of work in various departments of a printing plant and Number 590 is on the physiological cost of tailors' work. Curves are published showing very neatly the typical characteristics we have already noted in preceding paragraphs.

Other miscellaneous studies are Numbers 588, in which bicycle ergometry is shown to be more costly than stair-case work, and 599, in which the relative cost of swimming at different rates is determined.

Criticism and Defense of Waller Method. Hill and Campbell (291) criticize the Waller Method of determining energy cost in industrial work and cite an experiment in which contradictory results would be obtained by using the Waller Method instead of the complete Douglas-Haldane computation. They regard this discrepancy as evidence that the respiratory quotient must always be computed in work experiments and never assumed. These authors even go so far as to express skepticism over Waller's experimental finding that there is a steady rise in carbon dioxide output during continuous work. About the same time, Orr and Kinloch (453) added a further criticism of the neglect of the respiratory quotient in the Waller Method. Waller answered these two critical papers together, but we shall consider critical attacks by other experimenters before turning to his replies.

Gairns and O'Brien (206) published in the following year results of the only experiment ever performed in direct refutation of the validity of his method. These investigators ran several types of experiments and made a serious effort to check the various claims made by Waller concerning his technique. These authors evidently anticipated that their results would permanently and effectively stop all further experimentation with the method, and indeed they were right, since their work was published some months after Waller's death

and thus at a time when the method had no interested champion to defend it. The results obtained by Gairns and O'Brien certainly indicated a probability that the Waller Method cannot be as widely and as casually used as its originator thought possible, although we believe that the evidence which will be presented in succeeding sections of the present discussion will indicate that the situation is not so hopeless as these authors have supposed.

Greenwood and Newbold (232) are also severely critical of the method. They review important experiments utilizing the method and criticize especially Waller's advocacy of using carbon dioxide measurement alone. Their paper consists partly of a statistical treatment of Benedict and Cathcart's monograph which demonstrates the wide fluctuation of carbon dioxide output, and hence its low predictive value. These writers, moreover, even go on to demonstrate statistically that determinations based upon oxygen consumption or even upon oxygen and carbon dioxide together are but little more reliable than experiments based on carbon dioxide production alone. They conclude that "the practical conclusion seems, therefore, to be that, when any experimental calibration of different forms of muscular work is based on the confrontation of small sample (respiratory) measurements upon different subjects, only the roughest results are obtainable." Their rather involved arguments from regression equations, variation coefficients, etc., lead them to the conclusion that even under the best of present methods we can do little more than distinguish

between the energy costs of sedentary occupations and heavy manual labor. Pickworth (465), as we have already noted, found that the Douglas Bag Method yields basal metabolic rates which are too high by as much as 20 per cent, so if full dependence is to be placed upon the two studies last named we should be forced not merely to reject the Waller Method and other simplifications but would also be denied even the use of the standard Douglas bag technique, which has found world wide use in countless experiments.

Most of the criticism we have been considering has been based on statistical treatments of experiments in which oxygen consumption and carbon dioxide figures are available, so that results may be computed by the full method as well as by means of Waller's simplification. There are other criticisms of a more theoretical sort, particularly with regard to reliance upon carbon dioxide. The question is a live one in basal metabolic rate determination as well as in computing energy costs of work. The typical criticism of the use of carbon dioxide alone in respiratory exchange experimentation will be found in DuBois (172).

Any unbiased reader who goes through the various criticisms we have just mentioned will inescapably conclude that the Waller Method is utterly useless in any kind of experimentation and is even likely to doubt the value of any type of respiratory exchange determination even under the most refined techniques. Waller himself lived only to answer the early objections of Hill and Campbell, quoted above, and those of Orr and Kinloch. Three principal objections were raised by

these authors, each of which is discussed quite completely by Waller (595). The first objection was that one-half minute is too short a sample for reliable results, the second that the neglect of the food factor introduces serious error, and the third that the respiratory quotient must be determined rather than assumed. Waller makes answer to all of these points and shows also the advantages inherent in his method for actual study in industry. We shall not repeat his arguments here but shall add further evidence on these various points as gained from other authors.

Waller was able to give his method empirical confirmation in the above article by reports of minor experiments with parallel Waller and Douglas-Haldane computations. A study of marching (598) in which energy consumption was estimated by both methods demonstrated a reasonably accurate correlation between the Waller Method and the full Douglas-Haldane procedure.

It appears to the present reviewer that much of the criticism which has been directed toward the Waller Method would lose force if the claims for the method were made a little less pretentious. A considerable body of evidence has accumulated to support the view that the physiological adjustment to light work is an altogether different matter from the adjustment to heavy work, and that, while the Waller Method may be demonstrated to be unreliable for heavy work determinations, such demonstration does not invalidate its use in connection with continuous light work such as is ordinarily found in American industry. Modern

engineering has so reduced the heavy manual labor in American industry that we are not concerned so much with differences in energy cost of various methods of lifting, pulling, and pushing heavy loads as we are with differences in the adjustment of the work bench and factors of a like nature. It is useful, then, for us to consider separately the physiology of light and heavy muscular work, knowing that the Waller Method or any other technique which may be applied to light work loses little in utility even if it cannot be applied to heavy manual labor.

Hill (284), in a discussion on page 16, apparently confirms the feasibility of using carbon dioxide output when studying long periods of light work, and Viale (575 and 577) found the carbon dioxide discharge proportionate to the external work provided the task was light and continuous. He and Briggs (112) both speak of an individual maximum carbon dioxide discharge which is not exceeded no matter how greatly the external work is increased and Briggs even finds that "when exertion of steadily increasing magnitude is undertaken the expired carbon dioxide percentage first rises and then falls."

Other important physiological studies of the difference between light and heavy work are those of Hough (298) and Hill, Long, and Lupton (285). In this latter study again there are indications that carbon dioxide output may be used as an index of moderate exercise but not for heavy labor. Becker and Olsen (39) even find that (page 195) "in the case of minor muscular work, when the subject takes up the same

position during the entire experiment and works with unhindered respiration, the value of the increase of metabolism may approximately be expressed by the increase of the respiratory volume." This they hold to be true because the carbon dioxide percentage of alveolar air under such circumstances remains practically constant.

Sargent (497) demonstrates that the recovery rate following severe exercise is highly variable and also points out that there is an actual elevation in the basal metabolic rate following such exercise. Both these factors would tend to invalidate computations of energy cost of heavy muscular work by any of the respiratory exchange methods. The experiments of Campbell, Douglas, and Hobson (121) are useful in connection with the theory underlying the Waller Method and their results also appear to be negative to that method for use with brief periods of heavy work. In spite of these theoretical objections, however, experimental determinations of energy cost of heavy manual labor are being made on the basis of carbon dioxide production and are being reported in the journals. See, for example, the study by Gullichsen and Soisalon-Soininen (236).

Assuming that the references quoted in the last few paragraphs demonstrate the necessity of a separate treatment of the physiology of light and heavy muscular work and that respiratory exchange determinations adaptable to the one may be unsuited for the other, we may now consider separately the various objections raised to the Waller Method with the idea in mind that

its application will probably need to be restricted to light industrial work and should not embrace all types of manual labor as originally advocated.

Let us consider first the validity and reliability of taking a brief respiratory sample. Respiratory samples of much shorter duration than those employed by Waller were used in an experiment by Krogh (349) in 1913. "In the series of experiments on the initial stages of muscular work we found it necessary to make three distinct determinations of ventilation and respiratory exchange during the first minute of work. . . . The accuracy of such a respiration experiment lasting generally from three to six seconds is not great, certainly. The error may easily amount to ten or even twenty per cent, but we have found the method efficient for our purposes and no other appeared possible." He refers again to these experiments in his manual written in 1916 (354). "With violent exercise a bag taking 60 liters will not hold the air expired during one minute, but it has been shown (1913) that experiments of much shorter duration are sufficient to give perfectly reliable results." He asserts again (352) "that one can secure full and dependable estimation of an individual's resting metabolism with experiments of very short duration."

Bedale (41) contends that accurate determinations can be made from samples collected for less than one minute and, furthermore, that it is thus possible to obtain true samples of the working period. She determined oxygen consumption in her experiments, however, rather than carbon dioxide output as in the Waller

Method. Aitken and Clark-Kennedy (5, 6), in very recent investigations, have demonstrated the possibility of making reliable studies even of the various parts of a single respiratory cycle.

One of the alleged dangers of using short respiratory samples is the possibility of securing artificially high carbon dioxide output through over-ventilation or *Auspumpung*. King and Cross (323) conducted a careful investigation into the extent of such possible washing out of carbon dioxide and found that this can occur only once in successive experiments, so that if two determinations are made one can detect over-ventilation through failure of the two sets of data to check. This refers, of course, to the use of carbon dioxide determination in finding basal metabolic rates, but there should be little difficulty in working out similar check experiments when determining energy metabolism. It is Waller's contention that an experienced operator can detect *Auspumpung* through careful observation of the subject, and he asserts further that most subjects soon overcome any tendency toward forced breathing. Subjects who are unable to overcome this tendency would, of course, need to be excluded from experiments.

Henderson and Haggard (265) report that trained athletes do not blow off excess carbon dioxide even in maximum exertion or afterward, but that this is not true of untrained subjects. This fact should be of considerable significance since most of the subjects upon whom the Waller Method would be used would be likely to be in at least a more than average state of training.

King has done still other work (see 323 above) in support of the use of carbon dioxide in basal metabolic rate determination. Such studies have at least an indirect bearing upon the use of carbon dioxide figures alone in work experiments, and are thus worthy of notice here. In one study (322) he demonstrated from the figures of Benedict and Carpenter, and Soderstrom, Meyer, and DuBois that there is a somewhat higher coefficient of correlation between carbon dioxide and measured calories than between oxygen and measured calories. In his manual (321) there is a further defense of the carbon dioxide method, which he believes to be at least as accurate for basal rate determinations as determinations based upon oxygen consumption. McKinlay (421) also experimented with carbon dioxide methods in basal rate determinations, finding in 26 cases that the basal metabolic rates varied 3 per cent on the average from those computed by the full method and that in 33 other cases the oxygen consumption computed from carbon dioxide output differed by 4.8 per cent from the measured oxygen consumption of the same cases. The greatest fluctuation was found with diabetic patients and certain other pathological cases. Rabinowitch and Bazin (475) have also claimed eminently satisfactory clinical results from the use of carbon dioxide determination alone, and still further discussion of the usefulness of the carbon dioxide determinations in the clinic will be found in the contributions by Sainton and Peron (495).

Hendry, Carpenter, and Emmes (268), however, find some objection to reliance upon carbon dioxide in

the clinic and recommend the use of face mask and respiratory valve to reduce the tendency toward wide fluctuation found in certain subjects.

Turning now from studies bearing on the use of carbon dioxide determinations in the clinic, we find that such measurements have had a long and honorable history in experiments in working metabolism. Disregarding the very early experimentation on the part of Lavoisier and his successors, we may note that the early tradition of the Swedish school, for example Johansson (310), and Johansson and Koraen (311), in using carbon dioxide alone in the study of positive, negative, and static work, has been carried on down practically to the present time.

Another early study is that of Higley and Bowen (278) who found that "the output of carbon dioxide from the lungs is practically uniform from minute to minute during uniform muscular work, if the blood has had time to take part fully in the process of elimination." They mention further that upon cessation of work the output of carbon dioxide decreases to the normal amount in about two minutes but not until after a latent period of about 20 seconds. Since the samples taken under the Waller technique are collected within the first 20 to 30 seconds, their findings would seem to support the use of this method. Krogh (352) also approves of the use of carbon dioxide alone in Douglas bag experiments in which simplicity and saving of time are of greater importance than extreme accuracy.

Benedict and Johnson (77) used and described a large chamber for the studying of groups of thirty to

forty persons by means of carbon dioxide elimination alone. "The determination of the carbon dioxide production has of itself but little value, but is used for computing the heat production by means of the calorific value of carbon dioxide at an assumed respiratory quotient. Such a method is admittedly not so accurate as the direct measurement of the heat production or the computation of the heat production from measurements of the oxygen consumption. It is, however, a rapid method and not too costly for determining the approximate heat output of a group of people." Their measurements, taken two hours after a light breakfast, indicated clear separation between occupations involving various degrees of muscular activity. For example, there was a 3 per cent increase in metabolism while reading aloud as compared with silent reading and a 13 per cent increase in metabolism in light hand-sewing as compared with reading quietly. Boigey (95 and 97) is a well-known French physiologist who has successfully employed methods similar to the Waller Method in experiments on muscular exercise. He reports that "the variations of oxygen consumption and of carbon dioxide production are each proportional to the intensity of the work."

In 1922 (94) he wrote an enthusiastic account of the use of the Waller Method in appraising the energy cost of different types of athletic sports. Meters and bags are pictured and described, as is also the special carbon dioxide analysis apparatus in portable form, as devised and used by Waller. He agrees with Waller in discounting the effect of the respiratory quotient as

a matter of practice in the experiments for which the method was intended.

An experimental investigation of the relation between gaseous exchange and respiration in severe muscular work is reported by Herxheimer and Kost (272) who report that carbon dioxide production is proportional to the ventilation but that this is not true of the oxygen consumption.

The reader who has covered the material on the respiratory quotient, as outlined in the section under that heading, will realize that the modern conception of the respiratory quotient is somewhat different from that which prevailed at the time of the criticism of the Waller Method because of its neglect of the R.Q. More is known now about the ratio ordinarily to be expected, both as to its stability, and the range within which it may fluctuate. Along with this greater empirical knowledge has come a less dogmatic view of the theoretical significance of the ratio.

The subject is so complex that it is impossible to prove or disprove any point on the basis of one set of experimental data alone. For example, there are studies such as that of Benedict, Emmes, and Riche (74) indicating a rise in the R.Q. following a carbohydrate meal, against which may be set the assertion from Benedict and Murschhauser (79) that "the heat production per unit of work is practically independent of the taking of food." Again, Carthcart and Burnett (133) report a rise in the R.Q. during work, whereas Henderson and Haggard (265) find the working R.Q. to be the same as in the preceding rest state.

The immense amount of work by Hill (281) and his collaborators, as well as the evidence of Krogh and Lindhard (356, 357) and other experimenters, establishes a strong presumption that the working R.Q. is fairly close to unity. It seems safe to assume from these studies and from the recent investigations of Bock, van Caulaert, Dill, Fölling, and Hurxthal (91) that if the R.Q. is treated as constant at about 95 there will be little likelihood of introducing serious error through failure to compute the R.Q. Smith (525) found the majority of R.Q's. between 80 and 95 with very few in which the deviations were extreme.

There remain to be mentioned a few miscellaneous points both in favor of the method and in opposition to its use. Carpenter (124) found decreasing amounts of carbon dioxide during work periods, which was contrary to the typical findings of Waller, and Mague (552) contended that fatigue has little or no effect upon the respiratory phenomena caused by work, which is also contrary to Waller's experimental results. Polokov (468), however, found a fairly significant rise in carbon dioxide output with the continuation of muscular work. This confirmation of Waller's results is the more surprising on account of the impossibly bad technique used in Polokov's investigation.

A possible limitation is seen in the study of Hindmarsh (294), in which reliance upon carbon dioxide records alone tended to obscure the lowering of basal metabolic rates found in warm climates. It is possible that certain minor factors of industrial interest might also be thus obscured. A more definite limitation in

the use of the Waller technique is indicated in the study of Schneider and Clark (504), who found that carbon dioxide output is totally unreliable as a measure of work performed under low atmospheric pressures. Possible advantages of the technique, however, are seen in the studies of Griffith, *et al.* (233), who discovered the rather dubious advantage that carbon dioxide output is not subject to seasonal fluctuations but that oxygen consumption is, and of Benedict and Cathcart (72), indicating that there is not as long an after-effect from work on carbon dioxide output as there is on oxygen consumption. This would appear to mean that tests using carbon dioxide alone could be terminated earlier.

We have devoted considerable space to a consideration of the Waller Method because of its great potential significance in industrial work. The author of this review feels that he has collected sufficient evidence to entitle the method to a completely new evaluation, on the grounds that there is ample evidence either to refute or else severely limit nearly all of the charges made against the method, provided its application is limited to the ordinary range of industrial labor. The technique is of sufficient importance to demand that it shall not be overthrown on the basis of a single experiment, nor can it be established on such a basis. Similarly, we cannot expect either to destroy or to verify the method on purely theoretical grounds. There is too much to be said on each side.

There are two important research projects in connection with the method which demand immediate at-

tention. One of these is the conducting of several types of exploratory experiments by independent investigators, some of whom should perform experimentation in industry and some of whom should work in the laboratory. The establishment of the validity of a respiratory technique is too involved a problem to leave in the hands of a single research worker. The other type of work now needed is a pushing forward of the method of rechecking the statistics accumulated from other investigations, as has already been done to the detriment of the Waller Method by Greenwood and Newbold, quoted above, and to the advancement of the method by Waller himself and indirectly to its advancement by King. There is available a large literature which may be worked over, using both the Waller Method and the Douglas-Haldane computations in order to determine the limits in accuracy to be expected from the former. Since this project will yield results in a shorter space of time than would be necessary in laboratory experimentation, the writer urges that some worker with adequate statistical training review the studies of muscular work listed in our bibliography with this purpose in mind. Useful data will be found in early studies such as those of Douglas and Haldane (168) and Carpenter (124) and many other studies preceding these. Tabulations of oxygen consumption and carbon dioxide output will be found in many of the Carnegie Institution publications such as that of Benedict and Murschhauser (79). Even minor studies on pathology and special conditions, such as that of Peabody and Sturgis (461), will be found use-

ful in that separate carbon dioxide and oxygen data at rest and during exercise are supplied. The recent sources are fully as valuable, so the investigator should not overlook the studies by Bock, *et al.*, (91) which appear to give some support to the Waller Method, as well as taking into consideration the large amount of modern work done in Germany by such men as Simonson (519-522).

Gas Analysis Under the Waller Method. We have seen already the desirability of experimenting to establish the limits within which the Waller Method may be applied to industrial problems. Anticipating that investigators will wish to carry out such experimentation in industry with an apparatus as simple, portable, and accurate as possible, we have collected information on a variety of methods of gas analysis from which choice may be made to suit local conditions and needs.

One of the most promising techniques for the determination of the carbon dioxide content of expired air depends upon the difference in electrical conductivity of a platinum wire surrounded by a gas mixture in which the proportion of carbon dioxide is variable. Daynes (156), in a mathematical paper, discusses the theory of a device constructed on the principle originally developed by G. A. Shakespear. Another illustration and description of the device appeared anonymously (12) in a note concerning the apparatus manufactured by Cambridge and Paul Instrument Company, London. In the form described two platinum wire spirals are suspended in gas mixtures

and placed on opposite arms of a Wheatstone bridge. The carbon dioxide content of the gas surrounding one of the spirals may be varied and, by noting the deflection of the galvanometer, readings of carbon dioxide percentage may be directly made. This original device was intended for use in the flumes of boiler plants, but Hill (279), in 1922, sent a note to the Journal of Physiology to the effect that the instrument was being tested at the Physiological Laboratory experimentally and at the Royal Infirmary of Manchester clinically in order to determine whether the device could be adapted to physiological use in measuring the carbon dioxide content of respired gases.

The next important contribution to the method was made by Palmer and Weaver (459) who demonstrated the rapidity and accuracy with which simple gas mixtures with only one gas variable may be analyzed. These workers did not use a calibrated galvanometer but determined their carbon dioxide percentages in terms of resistance necessary to bring the galvanometer needle to zero. Using this type of apparatus they were able to attain readings which usually checked to the last decimal with Haldane determinations. About the same time Knipping (336) advocated modifying the Siemens and Halske instrument for determination of nitrous oxide in such a way as to make possible determination of carbon dioxide percentage in alveolar air. (The Siemens and Halske instrument is described by Palmer and Weaver, page 43.) Knipping (327) describes the modified Siemens and Halske device and advocates its use particularly in connection with

anaesthesia and other special clinical purposes. In the meantime Rabinowitch and Bazin (475) described their use of the thermal conductivity method in determining basal metabolic rates. They appear to have used the Cambridge Instrument Company's device. Remarkable correlations with basal metabolic rates determined by the Tissot-Haldane Method were reported.

Ledig and Lyman (370) have recently described the construction and operation of an apparatus of this type capable of recording carbon dioxide accurately to .01 per cent and oxygen to .02 per cent, thus satisfying all the needs of respiratory work. The apparatus is costly and delicate, however, and is made so principally on account of the oxygen analysis feature. A slide wire is used to bring the galvanometer to zero, as an improvement upon the direct reading galvanometer. Technical details of construction are given, such as methods of holding the electrical current constant and maintaining constant temperature around the gas cells.

Illustrations of electrical conductivity devices and descriptions of their use will be found in Knipping and Rona (343), pages 171-175. The Cambridge Instrument Company Catalog, No. 5 (119) describes their Portable Carbon Dioxide Indicator for Alveolar Air on page 15. This is developed from the early Shakespear type, in line with suggestions made by A. V. Hill. It is calibrated from 0 to 10 per cent, with an accuracy of one-tenth of one per cent. Electrical carbon dioxide meters are made for engineering use by the Brown Instrument Company (114) and by Leeds and Northrup (372). In a private communication, the Brown Instru-

ment Company asserts that a model can be made at reasonable cost with an indicating scale from 0 to 10 per cent, at an accuracy of one-half of one per cent. The carbon dioxide meters supplied by the Brown Instrument Company and by Leeds and Northrup can both be made continuously recording if desired. Willard and Smith (612) describe a compound commercially known as Dehydrite, which is used by Ledig and Lyman in their Katharometer.

Other Simplified Methods of Gas Analysis. Efforts to simplify gas analysis have taken many forms and analysis instruments of several widely different types are available. We have already considered the use of electrometric methods of determining carbon dioxide percentage, employing Shakespear's Katharometer in one form or another. An electrometric method of an entirely different sort is described by Bayliss (37). "The method is chiefly valuable for determining the amount of carbon dioxide absorbed by a caustic soda solution and it is believed that it forms the best and quickest method for rapidly absorbing and measuring relatively large quantities of this gas." Fenn (191) also describes a somewhat similar method based upon measurements of the conductivity of carbon dioxide dissolved in barium hydroxide. It is possible with this method to detect 1×10^{-7} of a gram of carbon dioxide. Applications of conductivity methods of determining carbon dioxide in solution are also discussed by Nicloux (449). These methods did not appear easily adaptable to industrial use, so no attempt has been made to cover the literature concerning them.

In a preceding section we mentioned certain of the simplifications which had been proposed for use in connection with some form of Haldane gas analysis. There still remain two or three other modifications which must receive our attention.

Boigey's account (94) of Waller's portable gas analysis apparatus for the measurement of carbon dioxide in gas mixtures has been quoted above. He has written another article (95) which appears to deal with similar topics, but which was unavailable to the reviewer.

Another very simple type of apparatus is described by Abady (1) on page 381 of his *Gas Analyst's Manual.* Oechelhauser's carbon dioxide apparatus is there described as a simple potassium hydroxide absorption arrangement which apparently can be read to about one-tenth of one per cent. There appears to be a possibility of serious error, however, in the lack of protection of the gas mixture from the potassium hydroxide *prior to taking the initial manometer* reading. Another drawback is the evident necessity for using large samples.

King (322) and Wardlaw (607*a*) are in favor of an open circuit method in which the carbon dioxide is collected in soda lime and its amount determined gravimetrically. Krogh (354) also states, page 43, that "when a gas analysis is considered a thing to be avoided the contents of the (Douglas) bag can be taken through a Haldane set of vessels for absorbing water vapor and carbon dioxide and the total carbon dioxide determined by weighing." The most evident drawback of gravimetric methods is the fact that a delicate and expensive balance is required.

Melka (423) describes a simple apparatus to analyze respiratory gases, but the journal in which this is described was unavailable to the reviewer. Bounhiol (108) describes an open circuit arrangement, known as a "Sphyximetre," which is intended to measure the oxygen consumption of the parts of a single breath. Two other interesting devices are those of Guillaume (235), who describes a spectroscopic method of gas analysis for clinical use, and Griffiths (234), who describes an elaborate and costly instrument for analyzing gases through utilizing the difference in the rate at which sound passes through them. The device is not likely to be of service in respiratory metabolism investigations, as the results are of a low order of accuracy, and the special features of the method are of little use in physiological experiments.

Other Open Circuit Methods of Metabolic Rate Determination. There are several types of open circuit methods of determining metabolic rate which have not yet been considered in these pages, as well as certain forms closely related to the Haldane methods which have already been described. Of the latter type is the old Tissot Method (562) and the majority of the modern methods described by Klein and Steuber (325). Here belongs also the account by Labbé and Stévenin (359) of basal metabolism determination, using the Laulanié gas analysis apparatus in conjunction with a Verdin spirometer, which is a dry gas meter calibrated decimally in liters. These same French gas analysis instruments are also described by Amar (10), along with the description of other French types of analysis apparatus.

An important historical document, but one not likely to interest the industrial worker, is the early monograph of Benedict (49) on "The Composition of the Atmosphere with Special Reference to Its Oxygen Content." There is a description of the Sondén-Petterson apparatus, which is an elaborate and sensitive development from the original Orsat principle. The chemistry of the method is discussed and other information, such as a history of air analysis, is supplied, as well as complete data on determinations made near the laboratory, over the ocean, and on Pikes Peak.

Other gas analysis methods for use under special conditions are those of Gappellen and Noyons (207) and Süffle, Hoffman, and Walz (549).

Turning now to open circuit methods in which the respiratory exchange is determined while the experiment is in progress, we find an interesting variety of ingenious arrangements for recording either the carbon dioxide production or the oxygen consumption or both.

Gesell and McGinty (213) constructed a very elaborate apparatus employing two different principles in determining respiratory exchange and in recording on smoked paper the variations in percentages of the two gases concerned. In determining percentage of carbon dioxide use was made of the fact that the acidity of a water solution varies with the partial carbon dioxide pressure with which it is in equilibrium. This difference in acidity was measured with a potentiometer by means of an ingenious recording arrangement. The oxygen was recorded continuously through utilizing

the fact that a Bunsen burner flame temperature varies with the percentage of oxygen present in the air supplied. The flame was made to play on a current of water, and records were made by use of a thermocouple vessel and a potentiometer. The apparatus as described was used successfully in animal experiments.

The description of the apparatus developed by McClenden, Anderson, Steggerda, Conklin, and Whitaker (415) has been abstracted in *Physiological Abstracts,* Volume 13, No. 2927: "An entirely self-contained spirometer is described which resembles in principle a Haldane gas analysis apparatus, in which the subject's lungs form the oxygen absorption chamber and a soda-lime tower absorbs the carbon dioxide. Of the two spirometer domes (each 100 liters), one is initially filled with carbon dioxide-free air and the other empty. The subject inspires from the former and expires into the latter by virtue of suitably arranged valves for 10 to 15 minutes. The air is then forced back through the soda-lime into the first spirometer. From the known volumes of air expired, and of the latter after removal of carbon dioxide, respiratory quotient and oxygen consumption rate are obtained. Constant temperature is maintained throughout by immersing the apparatus in a large bath of water. The volumes may be read to 10 cc., and buoyancy errors eliminated by an automatic chain device."

A standard open circuit method widely used in Germany is that which utilizes the Zuntz-Geppert apparatus. An early description of this apparatus was written by Magnus-Levy (398), and a recent description

of a modern form by Paechtner (457). Zuntz (623) describes a semi-portable form used on the continent, the apparatus being adapted for use in studies of work metabolism. It consists of a dry gas meter strapped to the back with a special arrangement for taking a series of minute samples and collecting them within a single sampling bottle.

Regelsberger (480) describes a complex electrical and chemical apparatus making it possible to record carbon dioxide automatically with an accuracy within one-tenth of one per cent.

One of the most interesting open circuit arrangements is that of Hanriot and Richet (247). This very early form of apparatus has also been described by Higley and Bowen (278), from whom our description was taken. Hanriot and Richet used an inspiratory gas meter, an expiratory gas meter, absorption apparatus for carbon dioxide, and, finally, a third gas meter. A moment's reflection will show that both carbon dioxide and oxygen volumes can be derived from the differences in the three volumes registered. Krogh (354) says of the method (page 46), "The experiments made by Hanriot and Richet are not particularly accurate, but in the writer's opinion there is no doubt that the possibilities of this method are great. With modern gas meters of sufficient size placed in one water bath, volumes can be measured accurately to 1/10,000, and arrangements could easily be made giving a continuous graphic record of ventilation, oxygen absorption and carbon dioxide output." In the light of modern knowledge it would also seem necessary to add a blower or

similar device to aid the lungs in driving the respired air through the three gas meters and absorption bottles.

Total Pulmonary Ventilation as an Index of Energy Consumption in Light Work. In our citation of various authors whose work seemed to support the Waller Method we quoted the assertion of Becker and Olsen (39) to the effect that pulmonary ventilation alone might suffice as an index of energy consumption, provided the determinations were made at a time when the body was in a state of physiological equilibrium. A similar point was made many years previously by Hanriot and Richet (247). These workers found very little fluctuation in carbon dioxide percentage when the ventilation in liters ranged from 11 to 18 per minute. They definitely recommend ventilation determination as an approximation method in light work experiments, although they mention that it is not suitable for use in heavy work.

No one, apparently, has made a careful study with the specific purpose of determining the validity of ventilation records alone as a substitute for more complete respiration experiments. The article by Kharina-Marinucci (320) was unavailable to the reviewer, so he does not know whether or not it may be a contribution of this sort. There have, however, been several studies in which the authors suggested the possibility of using total ventilation as an approximate index, and data are available in many other articles which may be used in checking the validity of the method as has been suggested in the case of the Waller Method.

Hansen (248) demonstrated a parallelism between carbon dioxide output, oxygen intake, and total ventilation in an experiment on the bicycle ergometer. Knoll (344) even gives consideration to the relationship between respiratory rate and the volume of air expired, although there seems to be little likelihood that respiratory rate can ever be developed into even an approximately reliable index. One of the best general treatments of total ventilation as an index of energy exchange is that of Magne (396). He discusses this subject in connection with his general account of respiratory exchange during muscular work and quotes considerable tabular data from other experimenters.

It must be recognized without question that many serious objections of a theoretical nature can be raised against the use of total ventilation as suggested, but there is evidence that it is at least more accurate than the use of loss of weight, and certain other factors which have been used in the past as an index of energy cost. All of the reasons which have been urged against the Waller Method may be put with even more force here, as well as a few special considerations applying only to this approximation method. The method, if developed for industrial use at all, could be used only to show relative differences, as there is but little likelihood that the results could be expressed in terms of calories as Waller does in his computations.

Boigey (97) has already expressed himself in opposition to the use of this method, saying that "the total ventilation is not always in accord with the quantity of oxygen consumed or carbon dioxide produced."

Herxheimer and Kost (272) show a close correspondence, in a recent experiment, between carbon dioxide output and total ventilation but a lesser correspondence between oxygen consumption and ventilation. This would be at least partly negative to the use of ventilation figures alone. The use of ventilation figures would also be contra-indicated under unusual atmospheric conditions and in the case of pathological subjects. McCann (413), for example, reports five cases of advanced pulmonary tuberculosis in which the ventilation rate is double the ordinary rate.

In keeping with our suggestion that the literature should be reviewed to discover the reliability and validity of total ventilation, we are enumerating a number of studies arranged chronologically which appear to be especially useful for this purpose. In addition to the studies listed for use in verifying the Waller Method, the following investigations should be found significant.

Durig (177, 179), two studies in 1911 in which ventilation, carbon dioxide, and oxygen data are supplied in full. Douglas and Haldane (168), a study in 1912 giving detailed results of an experiment on walking. Their oxygen and carbon dioxide as well as total ventilation figures may be taken to show the reliability of ventilation alone. This has been done by the present writer and one of the curves has been reproduced in a recent article (Page, 458). Liljestrand and Wollin (379), a study in 1913 supplying carbon dioxide tables and data concerning ventilation. Boothby (100), a study in 1915 seeming to confirm ventilation as an index.

Ilzhöfer (301), a careful experiment in 1925, with full details as to ventilation, oxygen, carbon dioxide, and respiratory quotient. Simonson (519-522), a series of recent experiments in which ventilation figures are shown to have some significance in industrial investigations. Bock, *et al.* (90, 91, 556), a very recent series of painstaking and carefully controlled experiments in which a variety of figures are supplied, among them being complete information on total ventilation.

Closed Circuit Methods for Determining Oxygen Consumption. An account of the theory and use of closed circuit methods of respiratory exchange determination will be found in nearly all of the various manuals which have been referred to in previous sections. Their importance is such, however, that we must refer specifically to several types of closed circuit apparatus and review the literature more specifically than is done is most of the general references available in English. The most prominent of the closed circuit types of apparatus in use is the Benedict form in one or another of its many varieties. It is worth our while here to review the history of the Benedict apparatus, tracing the many stages of its evolution down to the present time.

The old Benedict Universal Apparatus was described in detail by Benedict (47) in 1909. This is the original instrument from which the long series of modifications has developed. It was an elaborate and costly apparatus constructed so as to permit determination of carbon dioxide output and oxygen consumption. Then in 1916 Benedict and Tompkins (82) described the

Benedict Bed Calorimeter, a modification of the old Universal. The description of the calorimeter appeared in an article criticizing certain types of apparatus then current and enumerating several desiderata in metabolic apparatus and methods. These authors insisted upon accurate respiratory quotients and freedom from factors inducing *Auspumpung* and other disturbances of respiration. The Benedict Bed Calorimeter was said to embody these desirable features and to be sufficiently accurate for all clinical use.

Following this we come to the historic reference (51) in 1918, to the well-known Benedict Portable. While this device was intended to measure oxygen consumption alone, it is possible to determine carbon dioxide output by weighing the absorbing apparatus in a sensitive balance. The Benedict Portable was the first machine of its general type to employ a fan or blower. A second description of the apparatus and its method of use was written the following year (78). The next year Benedict and Collins (73) described the first simplification of the original Benedict Portable. The modification consisted of eliminating the three-way valve, the calcium chloride bottles, and certain minor parts and the assembling of the entire apparatus into a really portable unit. Complete instructions as to method of use were supplied in this article, as well as statistics demonstrating the validity of the apparatus.

In 1922 Roth (486-488) published a series of articles on a further simplification of the Benedict apparatus. He dispenses with the electric blower on the New Portable and substitutes flutter-valves of his own design.

He describes the use of the kymographic record in determining basal metabolic rates with apparatus of this kind, and supplies instructions for the use of his modified form. Evidence is given that his modification is as accurate as the old and possesses obvious merits in its own right as well. This same material, with the addition of certain charts and tables, was published in another source (485) the following year.

Sanborn, a manufacturer of metabolic apparatus, published information on the use of his form of the Benedict apparatus at about this same time. His book (496) comprises contributions from 21 workers in various metabolic fields, and constitutes a working manual for the use of the physician and technician.

The well-known Student Apparatus was described in 1923 (65), and a review (54) of all the types of apparatus used in the Nutrition Laboratory was written the following year. Chamber methods, portable apparatus, student apparatus, micro-respiration apparatus, as well as ergometers and miscellaneous equipment, were all described in this contribution. It also included an extensive treatment of respiratory exchange, heat production, and the various factors affecting metabolism.

Most of the references we have quoted so far are principally of historical interest to the industrial investigator, but in 1925 Benedict (58) wrote a description of a modification of the Benedict Student Apparatus which makes it possible to strap the device to a subject's back in order to study his working metabolism. In the same article there is a description of a

simplified and greatly improved Benedict-Collins apparatus. This device is only semi-portable but is useful where extreme accuracy is necessary. No investigator who intends to study working metabolism should overlook this description of the portable apparatus for work experiments, nor the theoretical discussion and comparison of open and closed circuit respiration apparatus which is also contained in this article. There is not much literature in English on the use of this new apparatus in industrial work, although the reader may be referred to an experiment by Benedict and Parmenter (80) in which this apparatus was employed. A French description (59) appeared in 1927 in which photographs are shown of the apparatus in actual use.

The latest modification described by Benedict (60) is that of a Field Respiration Apparatus for use in surveying racial differences in metabolism. It is a modification of the portable form described in reference 58 and "the apparatus has been so simplified that it enables the determination of the oxygen consumption (apparent volume) of an individual *with but one major measurement,* i.e., the time required for the absorption of six pumpfuls of oxygen."

The Benedict Apparatus has been widely used in Europe and has been subjected to various modifications there. We may mention the recent contribution of Müller (439) and the criticism and suggestions about handling the apparatus contributed by Dautrebande (153).

It must be noted, in closing our discussion of the Benedict Apparatus, that most of the older forms are now obsolete, and that Benedict himself has stated that

the Nutrition Laboratory has now definitely abandoned closed circuit chamber methods in favor of the highly sensitive gas analysis technique for use with an open circuit chamber developed by Carpenter (122) in 1924.

Another apparatus of considerable importance is the closed circuit wedge-spirometer, described originally by Krogh (349) in 1913. The apparatus has been given universal usage and was described ten years later in Danish (352), German (353), French (355), and English (350). A recent discussion of the method has been contributed by Stolz (540), who describes certain modifications and improvements. Periera (463) describes an experiment using the Krogh apparatus coupled with an open circuit arrangement for measuring the respiratory exchange by Haldane analysis. He succeeded in demonstrating that the oxygen consumption at rest while breathing air and while breathing nearly pure oxygen is substantially identical. Presumably this finding would also hold in the case of Benedict apparatus and other closed circuit methods.

A description of the Benedict and Krogh types by no means exhausts the list of closed circuit devices, as the following sampling will indicate.

Hannan and Lyman (246) describe a rather large and cumbrous apparatus similar to the wedge-type spirometer of Krogh's. The amount of oxygen consumed is obtained from the graph by plotting a straight line through the trend of the deepest expiration points and reading the fall in a given number of minutes. Another apparatus for measuring oxygen consumption graphically is the portable calorimeter described by

McClendon (414). Schadow (498) described a closed circuit apparatus without blower or valves in which a wide-bore tube leads from a mouth-piece to a spirometer within which is located the soda-lime. He reports comparison experiments with the Benedict Apparatus and the Benedict-Knipping apparatus in which there was very close agreement. It was not recommended, however, for use in certain pathological cases nor in work experiments. Herxheimer (271) describes a closed circuit apparatus so simple that ordinary laboratory tubing has replaced the usual wide-bore respiratory tubing, and the soda-lime is contained in an ordinary bottle with a rubber stopper. A mouth-piece and valves are used and there is no blower. Other simple types of apparatus are described by Jones (313) and by Soto and Torino (531). In the latter device the subject breathes through a mouth-piece back and forth into the spirometer.

Many supplementary devices and modifications of various parts of closed circuit apparatus have been proposed. A few such suggestions may be enumerated here.

Pierce (466) describes an integrating device for use with the Benedict Apparatus, eliminating the ratchet-wheel in connection with the spirometer. Guthrie (238) describes a convenient and accurate spirometer and also comments on absorption substances. "Different samples of soda-lime purchased in the open market show great variations in carbon dioxide absorption powers, some samples being several times more efficient than others." Roth (484) also comments on absorbents,

pointing out that a good carbon dioxide absorbent may have a low moisture absorbing power, which would interfere with the efficiency of the apparatus.

Moore (431) reports on a combination mouth- and nose-piece designed to take the place of the mouth-piece and nose-clip used on early forms of the Benedict Portable, and Hendry, Carpenter, and Emmes (268) show that the use of a mouth-piece with the Benedict Portable is superior to nose-piece or half-face masks. Knipping (338) asserts that the use of the kymograph is essential in all short-run experiments using the Benedict type of apparatus.

Closed Circuit Methods for Determination of Oxygen Consumption and Carbon Dioxide Output. In the preceding section we traced the evolution of closed circuit methods from the early form in which oxygen consumption and carbon dioxide output were both determined, giving the impression that development had been entirely in the direction of simplification, and that closed circuit methods with provision for the determination of both gases should be considered obsolete. There have, on the contrary, been many modifications of the closed circuit principle which have retained the original feature of determining both gases.

Although the Benedict forms now in use are constructed to measure oxygen consumption alone, the principal part of the many early studies of the Nutrition Laboratory were made with some form of the old Benedict Universal Apparatus (47). A few samples of the more important of these studies are the

well-known investigations of resting metabolism by Benedict and Carpenter (70), and numerous studies of working metabolism, such as those of Benedict and Murschhauser (79) and Smith (525).

One of the most elaborately constructed respiration calorimeters using the standard closed circuit method was that of the Russell Sage Institute of Pathology in Bellevue Hospital, described by Riche and Soderstrom (482).

A new form of apparatus in which both carbon dioxide and oxygen determinations are made has been developed by Simonson (517). The issue of the journal in which this apparatus was described was unavailable at the time of this writing, but it appears that the apparatus is a development of the Benedict form and is a portable instrument suitable for a variety of work experiments.

Probably the most important work which has been done in developing the closed circuit principle for the determination of both gases is the extensive experimentation on the part of Knipping in Germany. We have listed a rather large number of his writings in spite of the fact that there is considerable duplication among them, hoping that the wider selection of journals will make his material more generally accessible.

Articles 329, 330, 332, and 340 in our bibliography each supply fairly complete information concerning his apparatus. He reviews all of the important methods in use on the continent (332), concluding that the Benedict Portable is the most useful on account of its simplicity, and the lowered possibility of error

which this brings about. The Sanborn type usually supplied, however, does not yield carbon dioxide records and so it is impossible to compute respiratory quotients. Knipping then describes his modification of the Benedict apparatus in which the carbon dioxide is absorbed in a special flask so arranged that the gas may, after the close of the experiment, be driven out of combination with the potassium hydroxide and back into the spirometer bell. This is accomplished by adding a quantity of 15 per cent sulphuric acid to the potash solution in which the carbon dioxide has been absorbed. Further descriptions of the apparatus are found in an article (340) which emphasizes the desirability of determining both respiratory gases, the necessity of eliminating back pressure of any kind in the breathing system, and the need for an arrangement for circulating the air through the spirometer rather than depending on a direct connection and pumping it back and forth by means of the subject's own respiratory movements. He mentions that Krogh's apparatus provides for a circulation of air, but that it offers resistance to breathing and is liable to develop leaks. His own apparatus fulfills all three requirements, since both gases are measured. A blower is used to relieve back pressure, and the air is circulated as in the Benedict apparatus. He describes the simplicity of the apparatus and speaks also of the large amount of experimentation which entered into developmental work in overcoming technical difficulties with the use of the new flask. He claims ample accuracy for both carbon dioxide and oxygen and states that the apparatus is

adaptable either to long-time or short-run experiments on animals, children, or man.

A brief description (339) of the flask which he adds to the Benedict apparatus is available, as well as another description (333) in more detail. It is described also in an article (331) in which mention is made of the fact that the 47 per cent potash solution used is supplied by Merk, and that the flask is obtainable from Albert Dargatz, Hamburg 1, Pferdemarket. In another article (329) mention is made of the fact that the complete apparatus, consisting of pump, spirometer, and special flask, may be obtained from this same source. An article in 1925 (340) mentions that it is unnecessary to drive out the carbon dioxide by means of sulphuric acid for purposes of volumetric determination, but that its amount can be determined very accurately while still in solution by means of a hydrometer. There are further notes (328) on the use of the hydrometer or pycnometer in the determination of the specific gravity of the potash solution and hence the percentage of carbon dioxide absorbed therein. It is also possible to substitute a refractometer for the hydrometer in making this determination.

A later article (341) is particularly concerned with the mask to be used in connection with a closed circuit apparatus. Knipping favors a special type of full-face mask having an inflated rubber ring around the edge of a metallic cup. There is described in the same article a simplified form of the Knipping apparatus, as supplied to the Iceland Expedition.

Knipping (337) describes an alcohol-burning con-

trol device for use with his respiration apparatus and discusses the necessity of running control experiments with all apparatus of this kind. A complete account and description, with illustrations, of the various modifications proposed by Knipping, as well as a wide variety of other types of respiratory apparatus, will be found in the working manual on the technique of clinical gaseous metabolism determinations by Knipping and Kowitz (342).

Developments from the Krogh apparatus permitting the determination of both oxygen and carbon dioxide have been described by Liebesny and Schwarz (378) and Melli (424).

In a brief preliminary report (261) and a more comprehensive account (260), Helmreich and Wagner describe their difference-spirometer for the graphical determination of carbon dioxide and the respiratory quotient. Their apparatus is a novel and ingenious rearrangement of the usual closed circuit spirometer, making it possible to determine the respiratory quotient from the graphical record. Lehmann and Müller (375) describe an elaborate form of closed circuit apparatus for determining both respiratory gases. It is a complex and expensive apparatus only semi-portable.

Graphical Methods in Respiratory Exchange Determination. It should be obvious that many of the instruments, both of the open circuit type and the closed circuit type, are adaptable to graphic registration. In some cases the description of the apparatus indicates that continuous graphic records in a direct reading

form are produced and in other instances kymographic attachments are employed in recording respiratory exchange in a somewhat more indirect fashion. To enumerate the various devices in which graphic records are produced would, for the most part, be to relist a large number of the instruments already discussed in preceding sections. Such tabulation would include the continuous electro-metric methods of Gesell and McGinty (213) and the electrical and chemical apparatus of Regelsberger (480), as examples of true graphic methods, and the devices of Hannan and Lyman (246) and Liebesny and Schwarz (378), as examples of graphic modifications applicable to typical closed circuit methods.

Other closed circuit developments which we have not previously described are those of Hagedorn (240), Dethloff (158), and Burger and Dusser de Barenne (116, 180, 181). These will be described below.

Hagedorn has developed an elaborate apparatus and technique making it possible to record oxygen consumption and carbon dioxide output graphically. There are two gas meters mounted on a common shaft and operated as pumps by an electric motor; these connect to two spirometers each of which records graphically on a drum, and the entire assembly must be operated under water to insure identical temperatures in the meters and spirometers. One spirometer registers carbon dioxide, the other carbon dioxide and oxygen, from which the oxygen may be computed by difference. The arrangement proposed by Dethloff is a less complex arrangement which also records both

gases graphically, and Burger and Dusser de Barenne describe three forms of their apparatus as adapted to three different ranges of ventilation rates.

Graphically recording apparatus of still other types is described by Higley and Bowen (278) and Kleiber and Wirth (324). The former apparatus consists of a rather complex arrangement by means of which changes in the weight of the carbon dioxide absorbing apparatus, mounted on a balance, are recorded graphically by the lever arm of the balance. Excellent curves produced with this apparatus are published in their article. The device of Kleiber and Wirth is a complicated but highly ingenious arrangement of pipettes and manometers, constructed so that records are made on a photographic strip and may be translated later in terms of carbon dioxide percentage. Either provisions or corrections are made for constant temperature, pressure, etc., so that the apparatus is fully as accurate as the usual volumetric methods.

Breathing Apparatus for Respiratory Exchange Experiments. Several general manuals and comprehensive studies having to do with breathing apparatus in respiratory experiments have been mentioned in various preceding sections. Since the proper choice of the type of mask, mouth-piece, valve, or other minor part of respiratory apparatus is often of paramount importance, we shall, therefore, once more refer to the leading sources of information for assistance in making such choice, and shall refer as well to certain special studies which we have not previously mentioned.

A careful comparison of various types of valves, nose-

pieces, and similar apparatus, as well as a discussion of open and closed circuit methods, was written in 1915 by Carpenter (123), and in 1919 Hendry, Carpenter, and Emmes (268) published the results of their extensive investigations into every possible combination of open and closed circuit methods with mouth-pieces, nose-pieces, and half-face masks. Boothby and Sandiford (104) and Henderson (264) describe spirometers, valves, masks, and other equipment for open circuit methods, and Knipping and Rona (343) describe all such apparatus, as well as tread-mills, ergometers, and other accessory devices used in metabolism experiments.

Dautrebande and Davies (154) investigated the variations in respiratory exchange produced by wearing masks of different types. Newcomb (446), Rowe (490), and Knipping (341) describe masks especially adaptable for use in closed circuit arrangements. Bailey (27) speaks of the undesirability of mouth-pieces or nose-pieces and describes a full-size face mask for use with open circuit apparatus; and Rosenheim (484) speaks of the undesirability of mouth-pieces also, and proposes a fireman's mask which he found especially suited for use with young women. Krogh (349) mentions using Stent's substance with Bohr's mask and filling the dead space in the mask with plasticene. Moore (431) proposed a combination mouth- and nose-piece and Williams and Nolting (614) describe a mouth-piece constructed upon proper dental principles.

Bailey (29) describes an efficient air valve having

very low resistance, and valves especially intended for use with the Douglas bag have been described by Fulton (200) and Raper (476). A light and simple valve made by Soderstrom is described in detail by McCann (413).

Accessory Apparatus in Respiratory Exchange Experiments. The investigator in respiratory metabolism finds it necessary to select not only among various major types of apparatus and breathing appliances, but must also choose certain minor devices of an accessory nature. Mention has already been made in the preceding sections of information regarding sampling bottles, gas meters, carbon dioxide absorbents, ergometers, and such devices used in one way or another in respiratory exchange experiments. As a matter of convenience we are assembling here certain of the more useful references to these various types of apparatus.

Bicycle ergometers are described by Campbell, Douglas, and Hobson (121), Severinghaus (513), and Waller and DeDecker (588). A useful form of treadmill is described by Benedict, Miles, Roth, and Smith (78). Benedict (54) describes ergometers and related devices, and Lehmann (374) describes a special dynamometer and an arm ergostat.

The Verdin spirometer is described by Labbé and Stevenin (359), the Boullite spirometer by Dreyer (171), and a special spirometer by Guthrie (238). Amar (10) also describes the Verdin spirometer as well as the spirometer of Tissot and many other forms of accessory apparatus.

Gas meters are described by Boigey (94) and New-

comer (448). Hartwell and Tweedy (253) supply information concerning maximum, average and minimum ventilation rates to be expected in respiratory experiments, which should be useful in the selection of gas meters, Douglas bags, etc.

Sampling bottles are described by Bailey (27, 28), and McCann (413), on page 852, speaks of a special sampling bottle in which glycerine and saturated sodium chloride in equal parts are found to be as satisfactory as mercury as well as less expensive.

Soda-lime is favored as a carbon dioxide absorbent by Krogh (349), although Guthrie (238) speaks of the great variations in absorbing power of different samples of soda-lime obtainable commercially. Roth (484) tells of the difficulties of finding a carbon dioxide absorbent which also possesses a high moisture absorbing power. Willard and Smith (612) describe a highly efficient drying agent. Knipping (341) supplies a footnote description of the method of preparing the potash solution for use in his apparatus.

Benedict (55) describes a control device developed in the Nutrition Laboratory for testing all types of respiration apparatus for gas volumes, heat, valve leakage, etc. The apparatus is obtainable through W. E. Collins, 584 Huntington Ave., Boston. Knipping (337) describes a control device for use with his apparatus, and Seuffert and Nitschke (512) describe an experimental control for use with the Voit portable respiration apparatus. Goiffon (218) proposes an instrument intended to standardize and simplify the taking of samples of alveolar air.

A number of the studies published by the Nutrition Laboratory carry descriptions of special devices used in metabolism experiments, and should be referred to by workers interested in such special types of equipment. As a single example we shall mention the electrical pulse recorder and the electrical pneumograph described by Benedict, Miles, Roth, and Smith (78).

III

APPLICATIONS AND RESULTS

Uses of Respiratory Exchange Determination in Industry. The determinations have their widest applications in the comparative study of energy costs when one major variable in the environment or the method of work is altered, and the others held as nearly constant as possible. In this way, certain optimum values may be discovered with respect to the adjustment of working conditions and equipment; or the cost to the worker of various environmental conditions may be revealed. Studies toward determining the best adjustment of working conditions and equipment may be of two sorts: either an artificial situation set up in the laboratory for investigating certain generalized modes of working or types of equipment, the results to be widely applied industrially; or they may be actual field studies of a specialized industrial operation, usually of more limited application. The two avenues of approach will be considered separately below.

Laboratory Studies of the Components of Industrial Work. The various types of muscular work ordinarily encountered in industry have been analyzed into their essential components and the energy cost of each component has been determined through respiratory exchange determination. Generalized modes of working have also been investigated experimentally with a view toward determining optimum rates and most efficient arrangements of the mechanical system involved.

One of the commonest devices used in laboratory

studies of muscular work is some form of an ergometer. This apparatus is especially suitable for experiments in determining the effects of continuous work over periods of time and in the determination of optimum rates and loads. Benedict and Cathcart (72), using a bicycle ergometer, found indications of a loss of bodily efficiency at the higher rates of work. Furusawa (202) also found that the oxygen required to perform a given amount of work varies markedly with the speed of work, although he discovered that with constant speed and varying load the oxygen requirement rises as a linear function of the work done, so that there is no optimal load. Hansen (248), also using the bicycle ergometer, found decreasing pulmonary ventilation with increasing rate up to 50 to 60 revolutions per minute. The oxygen consumption and carbon dioxide output were also minimum at this rate. Optimum loads were determined as a by-product of the study by Bedale (41), and were more directly investigated in a later study along similar lines by Cathcart, Bedale, Macleod, Weatherhead, and Overton (132). This study tended to confirm Bedale's earlier conclusions. Bedale's experiment will be discussed in a later paragraph, as the primary concern was not with rates and loads, but with the methods of carrying. Optimum rates and loads were also investigated by Jaschwili (307), and Lindhard (382) used a known optimum rate on the bicycle ergometer in an experimental investigation of muscular work of short duration.

Other uses of the ergometer were made by Wang, Strouse, and Smith (607), who studied the effect of

obesity and undernutrition on muscular efficiency and fatigue, and by Viale (573, 575, 577) in a series of experiments on energy consumption during muscular work.

Benedict and Parmenter (80) and Waller and De-Decker (588) substitute stair climbing for the bicycle ergometer in work experiments.

The most important studies in which various types of work are analyzed into their muscular components have been carried on in Germany, principally by Atzler and his associates. In one study (15), he reports a detailed investigation of crank turning, using the Krogh apparatus, and a similar study of weight lifting, using the Zuntz-Geppert apparatus. Ten generalizations as to methods of work are presented as a result of his study. Some of these are obvious, such as the statement that one should use muscle groups adapted to the task, as locomotion by riding a bicycle is more efficient than by propelling a wheel-chair because the leg muscles are adequate to the task and the lighter arm muscles are not. Another rule is that through motion study we should not eliminate too many needless movements, as some of them may rest the heavily worked muscles and lightly exercise others. In fact, in light hand work, one should deliberately get up and walk around at times to stimulate the circulation. The work tempo should be alert, with rapid work and long rest pauses, but one should not speed up too greatly on heavy work. Static work and unnatural or strained body positions should be eliminated or reduced to minimum. Burdens should be carried so that the cen-

ter of gravity is vertically above the supporting part of the body. Clothes should be worn that do not interfere with movements of the body. Energy is wasted through wearing clothes that are too heavy. These experiments were reported originally in more detail by Atzler, Herbst, and Lehmann (24). Similar material will also be found in Atzler's contribution (17) to *Körper und Arbeit.* A much earlier German investigation along similar lines is the study by Reach (478) of the muscular work of rotating the handle of a milk separator.

Another type of investigation which has borne fruitful results in recent years is the experimental determination of energy costs of carrying loads by various methods. Bedale's study (41) was based on only one subject, but the technique of the experiment was carefully worked out and the resulting curves showed decided differences in the economy of weight carrying by different methods. Hewitt and Bedale (274) reported an investigation which was preliminary in nature to this one. Klingendahl and Pesonen (326) conducted a similar investigation except that two subjects were used and the experiment was carried on in a respiration chamber. Of the three methods studied, weight carrying on the back was most economical, carrying alternately in each hand was least economical, and divided between the two hands was between these two in economy. This literature was reviewed by Atzler and Herbst (22) who contributed also their own Douglas-Haldane study of energy consumption in weight carrying on a horizontal plane. These two

authors have also reported a study (25) of the economy of pushing and pulling loads on a horizontal surface. A number of useful optima were discovered. The article is reviewed in the *Journal of the American Medical Association,* Volume 88, page 1609 and in *Psychological Abstracts,* Volume 1, No. 1154.

Several rules of practice with respect to pushing and pulling movements of the arm were derived by Lehmann from his study (374) in which dynamometers and ergometers were used in conjunction with respiratory exchange determination with the Benedict apparatus. The physiological cost of the muscular movements involved in barrow work has been determined by Crowden (150), and the cost of the work of shovelling using shovels of varying heights and handle-lengths and with different loads has been investigated by Wenzig (610). As we have commented in an early section, one reason why studies of this general character have not been widely made in America will be found in our high proportion of mechanical handling of what would otherwise be heavy manual labor.

A type of study which has attracted a great deal of attention from Scandinavian investigators is the inquiry into the relationship between positive, negative, and static work. Johansson (310) and Johansson and Koraen (311) reported studies on this problem nearly thirty years ago. Johansson demonstrated an increase in metabolism in all three types of work, but showed a lack of correlation between sense of strain and actual metabolism. The question was worked upon later by Hammarsten (244), using Johansson's apparatus, and

by Frumerie (199), who found no relation between carbon dioxide output in static work and the subjective feeling of fatigue. A more recent experiment was performed by Cathcart and Stevenson (137), using the Douglas Bag Method in studying the arm extensors and flexors separately in the three types of work. The total metabolism during negative work was about three-fourths of that during positive, and static work had a metabolic cost of about four-fifths that of positive work. The subjective fatigue experienced was greatest in the static work, however. These authors qualified some of their findings through certain theoretical considerations.

The problem has been attacked more recently by Dusser de Barenne and Burger (182, 117). They surveyed the literature in the latter article and reported their own experiments in detail in the former. The earlier investigations were confirmed in that certain types of light static work were found to produce intense subjective fatigue without materially affecting the gaseous metabolism. The most recent of the Scandinavian work is that of Hansen, Hvorslev, and Lindhard (249), who find no difference in the force exerted in actively moving a weight and the static task of merely holding it.

Studies which appear to pertain to this general field but which were unavailable to the reviewer are those of Abramson (4), Boigey (98), and Efimov (183). The first of these articles is said to describe Johansson's apparatus for use in work experiments. The article by Efimov was referred to by Polakov (468) but has been

found to be unavailable in the large reference libraries of Chicago, New York, Washington, and Boston.

Energy Cost of Industrial Occupations. Several efforts have been made to determine the energy cost of various industrial jobs, although the interest has usually been in assessing the physiological cost per day in relation to food requirements rather than in determining differences in physiological cost of the same job performed under contrasting working conditions. Studies of the latter type which are of special significance to the industrial psychologist and engineer will be given separate treatment and we shall record here investigations having the former as their primary aim.

Moss (436) computed the energy expenditure in various types of work in a coal mine, using a special modification of the Douglas Bag Method. Greenwood, Hodson, and Tebb (231, abstracted in 230) determined the total metabolism of female munition workers, engaged at various tasks and under diverse working conditions. A similar study, carefully carried out and reported in detail, is the investigation of energy expenditure and food requirements of women workers contributed by Rosenheim (484). Various kinds of farm labor have been investigated by Farkas, Geldrich, and Szakáll (188), using the Douglas Bag Method, and a study of brick laying has been reported by Baader and Lehmann (26), using the Benedict apparatus.

A composite study of an interesting sort was reported a number of years ago by Amar, although the results have little practical significance today. Amar studied

the movements involved in using a file and determined mechanically the most suitable pressure to use, the best body posture to assume, and similar factors regarding this occupation. He also determined the energy cost of filing, although it should be noted that he employed a very long tube leading from his subjects to the gas meter, which would seem to be rather bad technique.

Those who are interested in learning the energy cost of various occupations may be referred to the work of the Royal Society Food (War) Committee (492), and to Chapter X of the book by Collis and Greenwood (141), as well as to a number of standard textbooks on physiology and nutrition. A summary of the Scandinavian work on determining the cost of industrial occupations in a respiration chamber, using carbon dioxide alone as an index, has been written by Becker and Hämäläinen (40).

Not much work has been done on the energy cost of the lighter industrial tasks, or clerical and other business occupations. Typewriting alone has been carefully studied. The reader may be referred to the work of Carpenter and Benedict (126) in 1909 and to the more recent work of Ilzhöfer (301), and of Schroetter (508). The latter appears to be a comprehensive study but was unavailable to the reviewer. It is abstracted in Giese's *Methoden der Wirtschaftspsychologie,* pages 501-502.

The energy expenditure involved in light household work was studied by Benedict and Johnson (77), using a respiration chamber employing carbon dioxide determinations alone. About twenty subjects were used,

several occupations were studied, and each experiment was repeated two or three times. Clear differentiation between tasks was shown. A more recent study, more elaborate in certain respects although employing only one subject, has been reported by Langworthy and Barott (363). This was abstracted in an article published the following year (364). It is of interest industrially to note the significant differences found in dishwashing at tables of various heights. A table 85 centimeters high results in an energy expenditure of 20 calories per hour above the resting requirement and the requirement is increased if the table is materially raised or lowered. At a height of 100 centimeters the requirement is 24 calories and at a height of 65 centimeters it becomes 30 calories per hour. This material and other information from these authors was quoted by Wheeler (611).

Several studies have appeared on the energy cost of various athletic sports. Swimming has been studied by Waller and DeDecker (599), who employed the Waller Method, and by Moog and Schwieder (430), who used loss of weight as an index. The latter experimenters found that swimming 1000 meters caused as great a loss in water from the lungs and skin as running 10,000 meters. Loewy and Knoll (384) studied the energy exchange in skiing, Geldrich (209) determined the energy consumption of horseback riding at various paces, Gullichsen and Soisalon-Soininen (236) studied the carbon dioxide excretion in fencing and wrestling, and duBois-Reymond and Peltret (99*a*) determined energy costs of muscular exercise and ten-

sion in gymnastic training. Oxygen consumption data for certain major sports are supplied by Knoll (344).

A few miscellaneous studies are the determinations made of the energy cost of organ playing by Farkas and Geldrich (187), the cost of playing ping pong by Blomberg and Johnson (86) and of miscellaneous minor tasks by Cathcart and Trafford (138). Cathcart and Orr (136) employed the Douglas-Haldane method in determining energy costs of various military duties, computed by the hour, day, and week. Bedale (42) made Douglas bag determinations of children's activities at work and at play in a private school.

Energy Cost of Walking. Such an imposing amount of work has been done on the physiological cost of walking and marching that it demands special treatment in a separate section. Studies have been carried on for the purpose of determining energy cost and food requirements, particularly of soldiers on the march, and on other phases of greater industrial significance such as optimum rate and effect of carrying loads.

The classic experiment on the physiology of walking is that of Zuntz and Schumburg (624) in 1901. This study has been criticized in the light of modern physiological knowledge but received wide recognition at one time. Another early investigation was Durig's study (177) in 1911. Durig determined the ventilation, carbon dioxide production, and oxygen consumption at various rates, finding some relationship between speed and cost but no evidence that slowing within reasonable limits increases costs. These and other early studies were reviewed by Benedict and Mursch-

hauser (79) as a part of their study in 1915. They reported detailed results on standing, sitting, walking, and running at various rates, with and without food. The food factor was found to have but slight relation to heat production per unit of work. This study was supplemented by another typically thorough Nutrition Laboratory investigation carried on by Smith (525). Smith also found that energy cost is not much affected by rate until a rather high speed is reached.

A distinction is usually made between the acts of walking and marching, although under service conditions it is often not clear that there is a physiological difference. Marching might be merely a special gait of walking or it might be walking under load or a combination of these two conditions. We shall notice a few of the experiments designated as studies of marching.

Waller has used his methods in several studies of the cost of marching under various conditions. In one study (586) interesting data on marching costs at various speeds will be found, and two others (597, 600) supply graphs of the energy cost of marching under service conditions, demonstrating a carbon dioxide production of one-tenth cubic centimeter per kilogram body weight per horizontal meter. A more elaborate study of marching, bayonet drill, and other military activities was made by Cathcart and Orr (136), and a special study of the optimum rate of marching was reported by Cathcart, Lothian, and Greenwood (134). They found an optimum rate of about three miles per hour, while Benedict and Parmenter (80), in a recent

experiment, found the optimal walking speed to be the somewhat lower rate of 65 meters per minute. These workers also found sauntering to be very uneconomical, although the accumulated evidence from various studies indicates that the optimum walking speed is not a sharply defined point, particularly at the lower part of the curve.

The studies in which the cost of carrying loads is made a part of the investigation of the energy cost of walking are of more interest industrially than most of the experiments we have been considering. Studer (548), using the portable form of the Krogh apparatus, investigated the metabolism of resting, standing, and weight carrying. Another study of various factors affecting the metabolism of walking is the recent investigation by Basler (35). Atzler and Herbst (25) studied the economy of walking and of pushing and pulling loads, rather than studying the cost of carrying them. The optimum walking speed was found to be only 51.4 meters per minute. Bedale's study of the energy consumption of carrying different loads (41) has been referred to in other connections. Atzler and Herbst (22) review the literature on this subject through 1927.

Studies of the energy cost of transporting loads must be based upon one or another of several possible base lines. Bedale investigated basal, ordinary resting, and walking without load, finding the energy consumption in the order of 31, 38, and 46, respectively. She found the second of these best suited to her purposes. The study of the physiology of standing by Simonson (520)

should be of interest to experimenters planning research on similar questions.

The physiology of running has been principally studied by Hill and his colleagues. We have selected the work of Furusawa, Hill, and Parkinson (205) and Hill and Lupton (287) as examples. In the former article, page 50, it is stated that "the fatigue is due to the enormous rate of expenditure of energy in running at top speed. One subject . . . . was developing 8½ horsepower at his maximum velocity and liberating more than four grams of lactic acid per second in his muscles."

The cost of walking, both in and out of training, has been reported upon by Waller and DeDecker (591), and the cost of walking on a slippery floor as compared with a smooth one has been investigated by Hietanen, Nikkinen, Nyyssölä, and Sternberg (276).

The Study of Industrial Working Conditions. Metabolic studies which have actually been made in industry for the purpose of determining the physiological cost of work under diverse working conditions are extremely few in number. In previous sections we have noted various studies which have been made in relation to such factors as the optimum load, the optimum speed, the optimum handle-length of tools, the most suitable methods of pushing and pulling and performing various tasks, the effect of changing the height of the work bench, and other related factors pertaining in some degree to industrial work. Many of these have not been worked upon with any degree of thoroughness and there is no need of our adding more

than has already been sketched in preceding sections.

It would seem that by adopting suitable physiological measures it should be possible to investigate a host of other industrial labor problems from the standpoint of metabolic cost and secure thereby information which would be unavailable through any other approach.

Several metabolic studies have been made of the rest pause, and these will be given consideration here, but there still remain a variety of questions which have scarcely been touched by these methods. It should be possible, through respiratory exchange determination, not only to adjust the number, duration, and spacing of rest pauses, but also to learn something about the proper length of the working day and the proportionate energy cost of overtime. Even the proportionate cost of night work and day work might be studied, as well as the physiological differences between the five-, six-, and seven-day working week. A little work has been done on the arrangement of material and the effect of special methods and equipment, but there are no metabolic studies of the effect of music and rhythmic sounds, of monotony and variety in tasks, or of the possible effect of differences in the aesthetic appeal of the workroom. One or two studies have been made of the effect of noise and distraction and of the effect of certain home conditions, such as loss of sleep, but more work needs to be done before practical conclusions can be reached. The effect of differences in illumination might be somewhat difficult to investigate through respiratory exchange techniques, and it is difficult to say whether these methods could be made to yield valuable information concerning the influence of differences in ventilation.

One important reason for the meager nature of the experimental results on these questions will be found in the limitations residing in the respiratory technique itself. It must be clear to any one who has pursued even a small portion of the readings recommended in this guide that ordinary routine respiratory exchange determinations are valueless in studies of mental work, and may be of negligible significance in studies of muscular work chiefly involving restricted muscle groups, or in which there is considerable muscular strain but relatively little activity. Many broad classes of industrial jobs and nearly all business occupations are thus excluded from investigation by these methods, although those industrial jobs which remain are still in the majority.

In suggesting the energy cost concept as a substitute for the current concepts of industrial fatigue it must be borne in mind that reference is made to industrial tasks involving principally muscular labor, although this labor need not be of a heavy type. Work involving muscular strain or visual strain, or work done in cramped or unusual positions produces subjective fatigue effects out of proportion to the energy cost involved and may have serious physiological consequences as well. The subjective fatigue concept is still valuable, and it becomes more useful if we restrict its use industrially to cases of this sort rather than use it as a blanket term covering several diverse and only partly related meanings. This will mean that such matters as visual fatigue and local muscular fatigue will need to be studied by introspective methods or by

such physiological tests as may be found suitable, or else reliance will need to be placed upon otherwise unsupported studies of the work curve taken in its broadest sense. It is possible that sufficient correlation will be found between general bodily tension and eye strain, for example, to establish measurements of the one as measures of the other, but there is not sufficient evidence to validate such a practice at the present time. Also it is possible that generalizations concerning working conditions which are found to hold for manual labor will be found by experience to be suitable in other instances, but this cannot be determined except upon a basis of experience.

The Effect of Noisy Working Conditions. The present writer was originally attracted to the field of respiratory exchange determination in industry in planning an experiment to measure the physiological cost of noisy surroundings in the factory and workroom. He found the planning of such an experiment to be a more complex task than was anticipated, with the result that the experiment has not yet been performed, but he has accumulated a number of references on the subject which will be given consideration here.

Only one serious effort has been made to measure the energy cost of industrial noise, an experiment by Laird, reported first by Dixon (164) and later by Laird (360). It was abstracted anonymously (11) and has been reported more recently in greater detail by Laird (360*a*). The Douglas Bag Method was used and laborious efforts were made to control extraneous physiological factors. Elaborate apparatus was used in recording the

amount of work being done on the typewriter and in standardizing the conditions of the experiment. In spite of the impressive description of the experimental controls employed and the precautions taken against possibility of error, the descriptions of the experiment do not reveal the most essential information of all, i.e., whether both respiratory gases were measured or whether reliance was placed on one of them alone. The indications are that only carbon dioxide production was measured; using the potash side of the Haldane apparatus. In the absence of protocols or tabular data, it is impossible to tell whether this is the case, but, if true, it would greatly vitiate the conclusions drawn. Since Laird reports as much as a 40 per cent increase in metabolism during work under noisy conditions as compared with quiet it would seem necessary to have complete data before accepting his generalized results.

The only other studies of the effect of noise are output studies such as that of Kornhauser's (346), general observations on the desirability of sound-proofing as in the article by Hannam (245), generalized observations on the clinical and medical aspects of the problem as written by Glibert (216), and arm-chair deductions concerning the physiological effects of noise such as those contributed by Hoogheimstra (296).

Several articles have been written by Spooner, in the first of which (534) he presented a classification of noise producers, which may have been useful at that date and which were copied without credit by Mowery (438). The only other material in Mowery's article consisted of either commonplace or else inaccurate

statements of fact, and the nature of Spooner's article may be inferred from the following quotation: "Without attempting minutely to describe the ear, which has been called the second gateway of wisdom and is regarded as the most complex of all parts of the human frame," etc., etc. Spooner describes Low's audiometer in a later article (536), which is duplicated in another reference (535). An article full of generalizations and pictures was written recently by Faulkner for *Factory* and this was abstracted anonymously (13) in the *Monthly Labor Review.*

Morgan (432) has contributed a splendid laboratory investigation of the effect of distracting noises and is now engaged upon an electrocardiograph study which was given a preliminary report by Snook (528). Both of the methods used by Morgan are useful techniques in studying problems of this sort in the laboratory, but they are beyond the scope of a paper primarily concerned with respiratory exchange determinations in industry and so will not be treated here in detail.

Rest Pauses. About the only metabolic work of industrial importance which has not been noted at some length in earlier sections of the present study is the experimentation which has been done with respect to the rest pause. The influence of rest pauses has been studied in several experiments, although a great deal of work remains to be done in connection with this important problem.

A good modern summary of metabolic and other studies of the rest pause has been contributed to the literature by Graf (224). Simonson (516) reported

a lecture on fatigue and recovery after physical work in which he discussed the study of the metabolic effects of rest pauses, holidays, and vacations by means of determining lactic acid production and gaseous exchange. Strauss and Bandman (547) speak favorably of metabolic studies of the effect of rest pauses, commending in particular that of Hill and his colleagues in the recovery period. Messerle (426) reviewed the literature and reported his electrocardiograph experiment in work and rest pauses. Shepard (515) used loss of weight to discover effects of rest pauses in work, finding that rest pauses have more of an influence on such loss than does output. Many of the curves published by Waller have an important bearing on the physiology of the rest pause, as he himself has stated and as has been pointed out by Page (458). In view of the fact that Waller's experimental results in this connection have been subjected to criticism, it is interesting to note that Benedict and Cathcart (72) found a drop in efficiency after the third hour of muscular exercise. These experimenters found that oxygen consumption increased proportionately more or less as Waller had reported a proportionate increase in carbon dioxide output, although the rise in oxygen consumption was not so pronounced.

This concludes our catalogue of original investigations of a metabolic nature which have a bearing on the adjustment of working conditions to the worker in industry. There is a vast field still untouched, although new experiments are constantly being reported. One who is interested in this literature should refer not only

to the physiological sources but should also investigate the engineering references from the time of Taylor (557) down through the modern handbooks such as that written by Dana (152), as well as material appearing in the engineering journals. Similarly, access should be had to the psychological periodicals and to psychological texts and reference books such as those of Giese (214), Poffenberger (467), and Weber (609). There is not much material on metabolic studies in industry in either of the German references mentioned, although such books often carry material which is at least of incidental interest. Poffenberger's text carries more material of this type than is usually found in textbooks on applied psychology.

Highly pertinent material is often contained in reports of technical experiments in psychology, as the discussion of the fatigue concept in Crawley's study (148) and the theoretical considerations on interpreting the work curve contributed by Farmer (189) and Spencer (533*a*).

Metabolic Tests of Physical Efficiency. There is one other possible use to which respiratory exchange determinations may be put in industry. Benedict (61) is inclining toward the belief that basal metabolic rate may be of significance as an indicator of the relative physical fitness, the level of vital activity, or the physical powers of an individual. A low basal metabolism seems to indicate a lowered physical fitness or vigor as compared with the vitality when a higher metabolic rate is found.

A still more obvious direction of industrial research

would be the duplication in a modified routine form of experiments resembling those of Benedict and Cathcart (72) on the efficiency of the human body as a machine. Indeed, this has already been suggested by Atzler (15) who uses respiratory tests as a means of selecting individuals for muscular work on the basis of their physical efficiency. Briggs (112, 113) speaks of the effect of athletic training in economizing pulmonary ventilation by utilizing a higher percentage of oxygen from the air, and excreting a higher percentage of carbon dioxide. He states that the normal percentage may even be doubled in the case of well-trained subjects and indicates that this may be a possible measure of physical fitness.

These measures do not appear to have been widely adopted for there is almost no material upon them in the books by Dreyer (171) or Bovard and Cozens (109), nor in the review by Martin (405). There is a single quotation from Schneider in the book by Bovard and Cozens: "Respiration tests are unpopular because they require a highly trained gas analyst and because most of us alter our rate and depth of breathing when we are being watched."

The first part of Schneider's assertion is, of course, untrue, and there is reason to believe that the difficulty he mentions in the second part of the quotation is one which can be overcome by using suitable precautions. Schneider's own physical test (500) is based upon the effect of exercise upon pulse rate and blood pressure. The rating it provides should be useful in industry although the test was originally developed for use with aviators. Scott (510) commends Schneider's point-scale rating and supplies additional data.

BIBLIOGRAPHIC NOTES

It is anticipated that this guide to the literature on respiratory exchange measurement will be used by many who are relatively unfamiliar with the sources of physiological, psychological, and medical literature. Such persons may need assistance in extending the references in the special field of their proposed inquiry, as well as in bringing their references down to date. For this reason we are repeating the titles of the general reference books of most value in a bibliographic way, as well as listing the periodical indices and abstract journals available in fields related to the subject matter of the various sections of this book. It is to be understood that the references in the bibliography are not complete or exhaustive in number, but it should be possible to amplify the bibliographic materials for any section with the aid of the sources given below.

Fatigue

Durig (176), about 200 references on industrial fatigue.

Durig (178), nearly 500 titles on theory of fatigue and its measurement.

Spaeth (533), several hundred titles on "The Problem of Fatigue."

Collis and Greenwood (141), brief bibliographies on fatigue and related subjects.

Frumerie (199), early literature on relation between carbon dioxide output and fatigue due to static work; 41 titles.

International Labor Office (302), bibliography on visual fatigue and protection of eyesight.

Bovard and Cozens (109), bibliography on physical tests; some useful references in relation to fatigue.

General Physiology

Bayliss (38), pp. 741-846. Titles classified by author but not by subject.

Morse (435), general references, pp. 23-26. Also chapter bibliographies.

Abstract of the Literature of Industrial Hygiene (626), Volume 1–, 1919–. (Published as part of the *Journal of Industrial Hygiene.*)

American Journal of Physiology Index (627), Volume 1, 1898-1912; Volume 2, 1912-1922.

Arbeitsphysiologie. Zeitschrift für die Physiologie des Menschen bei Arbeit und Sport (628), Volume 1–, 1928–.

Respiratory Physiology

Krogh (354), well-organized bibliography to 1916.

Gesell (212), 141 titles on regulation of respiration.

Knoll (344), 42 titles on respiration in sports.

Means (422), bibliography on dyspnoea and related topics.

Basal Metabolism

Boothby and Sandiford (102), remarkably complete bibliography of about 700 titles, arranged chiefly by institution of origin.

King (321), 300 titles, recent and valuable.

Boothby and Sandiford (105), 28 titles, principally on surface area.

Du Bois (172), numerous footnote references on all phases.

Grafe (225), extensive bibliographies on every phase of human metabolism.

Krauss (348), 114 titles on gaseous metabolism.

Benedict (56), 15 references of a general nature.

McCann (412), 304 titles on calorimetry in medicine.

Murlin (442), 209 titles on metabolism in infancy and childhood.

Talbott (555), 169 titles on basal metabolism in children.

Physiology of Muscular Exercise

Hill (281), extensive references up to 1924.

Hill (284), bibliography continued from 1924 to 1927.

Hill (283), selected bibliography of 21 titles on metabolism of work.

Hill, Long, and Lupton (285), bibliography on muscle physiology.

Hill and Lupton (286), references on muscular exercise and muscle physiology.

Briggs (112), brief bibliography on physical exertion and respiration.

Barr, Himwich, and Green (34), good bibliography on blood chemistry of muscular exercise.

McCurdy (418), good bibliography on physiology of exercise.

Bainbridge (30), French, German, and English titles on general physiology of muscular work.

Walther (605), French bibliography on work physiology.

Campbell, Douglas, and Hobson (121), references on the respiratory effects of work.

Cathcart (130), 77 references on muscle work and protein metabolism.

Cathcart, Lothian, and Greenwood (134), bibliography and summary of early work on walking and marching.

Cathcart and Orr (136), bibliography on marching and related work.

Miscellaneous Factors Affecting Metabolism

Møller (428), 150 titles in bibliography of general nature.

Morgulis (433), 1000 references on fasting and related topics.

Cathcart and Burnett (133), 33 titles on effect of diet.

Sundstroem (551), 194 references on the effect of climate.

Landis (362), 25 references on the effect of emotion.

Zeigler and Levine (622), 20 titles on effect of emotion on metabolism.

Mental Work

Grafe (225), 23 titles on mental work, as well as good bibliography on emotional states and mental disorders.

Day (157), good bibliography on mental work.

Benedict and Carpenter (69), review of early literature with bibliography.

Gillespie (215), good bibliography on influence of mental and muscular work on blood pressure and pulse rate.

Industrial Psychology and Management

Rossi and Rossi (649), comprehensive bibliography on personnel administration, including titles on fatigue, working conditions, and related topics.

Cannons (635), all phases of industrial efficiency and factory management, but nothing on energy consumption because of early date.

Beresford (629), lists only 9 books and about 40 periodical references on industrial psychology.

Berg (630), selective bibliography of English titles on management. A few references on fatigue and other personnel problems.

Society of Industrial Engineers (633), 23-page bibliography on

industrial engineering and management; published too early to include material on metabolic studies.

International Labor Office (632), unavailable to reviewer.

General Abstracts, Reviews, and Indexes

Index Medicus. First Series (637), a monthly periodical published 1879-1899, indexing general physiological and medical literature..

Index Medicus. Second Series (638), a monthly periodical published 1903-1920; essentially a continuation of Series One.

Index Medicus. Third Series (639), a quarterly index to medical and physiological literature, published 1921-1927. Covers world literature on medical subjects in widest sense.

Quarterly Cumulative Index (647), published quarterly 1916-1926. Not quite as comprehensive as the *Index Medicus,* Third Series, and having the further drawback that foreign titles appear in translation rather than in original language.

Quarterly Cumulative Index Medicus (648), began publication in 1927 as merger of *Index Medicus* and *Quarterly Cumulative Index.* Complete index to the world literature on medical, physiological, biological, and related subjects.

Index Catalogue of the Library of the Surgeon-General's Office, U. S. Army (636), began publication in 1880; now in the Third Series. Third Series, when complete, will constitute practically complete world index.

Berichte über die gesamte Physiologie und experimentelle Pharmakologie (631), Volume 1–, 1920–.

Jahresbericht über die gesamte Physiologie und experimentelle Pharmakologie mit volständiger Bibliographie (640), Annual, Volume 1–, 1920–.

Journal of the National Institute of Industrial Psychology (641), carries reviews and abstracts, Volume 1–, 1922-1923–.

Physiological Abstracts (642), a monthly periodical, starting publication in 1916.

Physiological Reviews (643), a quarterly publication carrying reviews of physiological subjects, starting publication in 1921.

Psychological Index (646), an annual index to world psychological literature, starting publication 1894.

Psychological Abstracts (644), monthly journal carrying abstracts of selected world literature, beginning publication 1927.

Psychological Bulletin (645), carries abstracts and reviews of psychological literature, starting publication 1904.

Bulletin of the Public Affairs Information Service (634), carries an index to literature on fatigue, working conditions, and related topics, Volume 1–, 1915–.

Abstract of the Literature of Industrial Hygiene (626), Vol. 1–, 1919–. Appears as a part of *The Journal of Industrial Hygiene,* Harvard School of Public Health.

BIBLIOGRAPHY

1. ABADY, J. Gas analyst's manual. New York: Spon & Chamberlain, 1902. Pp. 561.
2. ABDERHALDEN, E. [Ed.] Gasstoffwechsel und Calorimetrie. *Handb. d. biol. Arbeitsmeth.*, 1926, Abt. 4, Teil 10, 819-936.
3. ———. [Ed.] Methoden zur quantitativen Bestimmung des Stoffwechsels des Gesamtorganismus von Organen und Zellen. *Handb. d. biol. Arbeitsmeth.*, 1925, Abt. 4, Teil 9, (5 Hefte). Pp. ix+930.
4. ABRAMSON, E. Energieumsatz bei Muskelarbeit. *Arbeitsphysiol.*, 1929, 1, 480-502.
5. AITKEN, R. S., & CLARK-KENNEDY, A. E. The concentration of CO_2 in successive portions of an expired breath. *J. Physiol.*, 1927, 64, xvii-xviii.
6. ———. On the fluctuation in the composition of the alveolar air during the respiratory cycle in muscular exercise. *J. Physiol.*, 1928, 65, 389-411.
7. AMAR, J. Le quotient respiratoire. *C. r. Acad. des sci.*, 1928, 186, 166-168.
8. ———. Recherches mécaniques et physiologiques sur le travail du limeur. *J. de physiol.*, 1913, 15, 62-74.
9. ———. Organization physiologique du travail. Paris: Trans. as *The physiology of industrial organization and the re-employment of the disabled.* London: 1918. Pp. 371.
10. ———. Le moteur humaine et les bases scientifiques du travail professionnel. Paris: 1914. Trans. as *The human motor.* New York: Dutton, 1920.

10*a*. American Society of Mechanical Engineers. Discussion. Carbon dioxide as an index of fatigue. *Mech. Eng.*, 1926, 48, Pt. 1, 494—.

11. [Anon.] Physical and mental effects of noise. *Mo. Labor Rev.*, 1927, 24, 947-949.
12. [Anon.] A new carbon dioxide recorder. *Electrician*, 1929, 89, 15—.
13. [Anon.] The toll of industrial noise. *Mo. Labor Rev.*, 1927, 25, 56—.
14. ARBORELIUS, M., & LILJESTRAND, G. Muskelarbeit und Blutreaktion. *Skand. Arch. f. Physiol.*, 1923, 44, 215-236.

15. Atzler, E. Aufgaben und Probleme der Arbeitsphysiologie. *Dtsch. med. Woch.*, 1924, 50, 1280-1282, 1320-1323.

16. ———. Methoden der Arbeitsphysiologie unter besonderer Berücksichtigung des Problems der gewerblichen Ermüdung. *Handb. d. biol. Arbeitsmeth.*, 1928, Abt. 5, Teil 5a, H. 3, Lfg. 235, 427-558.

17. ———. Physiologische Rationalisierung. In *Körper und Arbeit,* ed. by E. Atzler. Leipzig: Thieme, 1927. Pp. 409-487.

18. ———. Stoff- und Kraftwechsel. In *Körper und Arbeit,* ed. by E. Atzler. Leipzig: Thieme, 1927. Pp. 147-194.

19. ———. [Ed.] Körper und Arbeit. (*Handb. d. Arbeits-physiol.*) Leipzig: Thieme, 1927. Pp. 770.

20. ———. Arbeitsphysiologie. Übersichtsreferat. *Jahresber. ü. d. ges. Physiol.,* 1924, 5, Pt. 1, 251-265.

21. Atzler, E., Betke, H., Lehmann, G., Sachsenberg, E., *et al.* Arbeit und Ermüdung. *Zentbl. f. Gewerbehygiene u. Unfallverhütung,* 1927, *Beiheft* 7. Pp. 91.

22. Atzler, E., & Herbst, R. Die Ökonomie des Lasttragens über eine ebene Strecke. *Arbeitsphysiol.,* 1928, 1, 54-74.

23. Atzler, E., Herbst, R., & Lehmann, G. Arbeitsphysiologische Studien mit dem Respirationsapparat. *Biochem. Zsch.,* 1923, 143, 10-46.

24. Atzler, E., Herbst, R., Lehmann, G., & Müller, E. Arbeitsphysiologische Studien II. *Arch. f. d. ges. Physiol.,* 1925, 208, 184-244.

25. Atzler, E., & Herbst, R. Arbeitsphysiologische Studien III. *Arch. f. d. ges. Physiol., 1927, 215, 291-328.*

26. Baader, E., & Lehmann, G. Über die Ökonomie der Maurerarbeit. *Arbeitsphysiol.,* 1928, 1, 40-53.

27. Bailey, C. V. An adaptation of the French gas mask for use in respiratory work. *J. Biol. Chem.,* 1921, 47, 277—.

28. ———. Apparatus used in estimation of basal metabolism. *J. Lab. & Clin. Med.,* 1921, 6, 657—.

29. ———. A low resistance air valve. *Proc. Soc. Exper. Biol. & Med.,* 1926, 24, 184—.

30. Bainbridge, F. A. The physiology of muscular exercise. (2nd ed.) (Monog. on Physiol.) London: Longmans, Green, 1923.

31. Bald, F. W. The regulation of respiration. *Amer. J. Physiol.,* 1927, 81, 222—.

32. Barbour, H. G. The effects of certain drugs and poisons upon the metabolism. *Endocrinol. & Metabolism*, 1922, 3, 717—.

33. Barcroft, J. The respiratory function of the blood. (Rev. ed.) London: Cambridge Univ. Press, 1925.

34. Barr, D. P., Himwich, H. E., & Green, R. P. Studies in the physiology of muscular exercise. *J. Biol. Chem.*, 1923, 55, 495—, 525—, 539—.

35. Basler, A. Über die gewöhnliche Schrittlänge und Geschwindigkeit bei einigen Formen des Gehens. *Arbeitsphysiol.*, 1928, 1, 271-277.

36. Bauer, V., & Blunt, K. Effect of a small breakfast on the energy metabolism of children. *J. Biol. Chem.*, 1924, 59, 77—.

37. Bayliss, L. E. Conductivity method for determination of carbon dioxide. *Biochem. J.*, 1927, 21, 662—.

38. Bayliss, W. M. Principles of general physiology. (Rev. ed.) London: Longmans, Green, 1924.

39. Becker, F. C., & Olsen, O. Metabolism during mental work. *Skand. Arch. f. Physiol.*, 1914, 31, 81—.

40. Becker, G., & Hämäläinen, J. W. Untersuchung über die Kohlensäureabgabe bei gewerblicher Arbeit. *Skand. Arch. f. Physiol.*, 1914, 31, 198-240.

41. Bedale, E. M. Comparison of the energy consumption of a woman carrying different loads, etc. *Indus. Fatigue Res. Board Rep.*, 1924, No. 29.

42. ———. Energy expenditure and food requirements of children at school. *Proc. Roy. Soc. London*, 1923, 94B, 368—.

43. ———. Two contributons to the experimental study of the menstrual cycle: Part II. Its relation to general functional efficiency. *Indus. Fatigue Res. Board Rep.*, 1928, No. 45.

44. Benczur, J., & Berger, I. Ueber die Wirkung thermaler Eingriffe auf die alveolare Kohlensäuretension. *Zsch. f. klin. Med.*, 1926, 103, 86-91.

45. Benedict, C. G., & Benedict, F. G. A permissible breakfast prior to basal metabolism measurements. *Boston Med. & Surg. J.*, 1923, 188, 849-851.

46. Benedict, C. G., Benedict, F. B., & DuBois, E. F. Métabolisme de l'homme séjournant dans un bain d'air chaud. *Bull. Soc. scient. d'hygiène aliment.*, 1925, 13, 444-446.

47. BENEDICT, F. G. An apparatus for studying the respiratory exchange. *Amer. J. Physiol.*, 1909, **24**, 345—.

48. ———. The influence of mental and muscular work on nutritive processes. *Amer. Phil. Soc. Proc.*, 1910, **49**, 145-163.

49. ———. The composition of the atmosphere with special reference to its oxygen content. *Carnegie Instit. Wash. Publ.*, 1912, No. 166.

50. ———. Factors affecting basal metabolism. *J. Biol. Chem.*, 1915, 20, 263—.

51. ———. A portable respiration apparatus for clinical use. *Boston Med. & Surg. J.*, 1918, 178, 667—.

52. ———. The measurement and standards of basal metabolism. *J. Amer. Med. Asso.*, 1921, 77, 247-250.

53. ———. Metabolism during starvation and undernutrition. *N. Y. Med. J. & Rec.*, 1922, **115**, 249-256.

54. ———. Methoden zur Bestimmung des Gaswechsels bei Tieren und Menschen. *Handb. d. biol. Arbeitsmeth.*, 1924, Abt. 4, Teil 10, 415-674.

55. ———. The control of gaseous metabolism apparatus. *Boston Med. & Surg. J.*, 1925, **193**, 583—.

56. ———. The measurement and significance of basal metabolism. *Lectures on nutriton, Mayo foundaton lectures*, 1925. Pp. 117—.

57. ———. The racial factor in metabolism. *Proc. Nat. Acad. Sci.*, 1925, 11, 342—.

58. ———. Some techniques for measurement of the gaseous metabolism of humans. *Boston Med. & Surg. J.*, 1925, 193, Pt. 2, 807—.

59. ———. Études récentes sur le métabolisme humain et animal. *Bull. Soc. scient. d'hygiène aliment.*, 1927, **15**, No. 4.

60. ———. A 'field respiration apparatus' for a medical and physiological survey of racial metabolism. *Boston Med. & Surg. J.*, 1927, **179**, 1161—.

61. ———. Age and basal metabolism of adults. *Amer. J. Physiol.*, 1928, **85**, 650-664.

62. ———. Basal metabolism data on normal men and women (Series II) with some considerations on the use of prediction standards. *Amer. J. Physiol.*, 1928, **85**, 607-620.

63. ———. Basal metabolism; modern measure of vital tivity. *Scient. Mo.*, 1928, 27, 5-27.

64. Benedict, F. G., & Benedict, C. G. Body posture minor muscular movements as affecting heat poducti *Proc. Nat. Acad. Sci.*, 1924, 10, No. 12.

65. ———. Ein einfacher Respirationsapparat. *Skand. Ar f. Physiol.*, 1923, 44, 87-102.

66. ———. Influence du milieu thermique ambiant sur métabolisme basal. *Bull. Soc. scient. d'hygiène alime*, 1924, 12, 541-575.

67. ———. Influence sur le métabolisme de la position corps et des moindres mouvements musculaires. *Bull. S scient. d'hygiène aliment.*, 1924, 12, 480-506.

68. Benedict, F. G., & Carpenter, T. M. Food ingestion energy transformations. *Carnegie Instit. Wash. Pu* 1918, No. 261.

69. ———. The influence of muscular and mental work metabolism. *U. S. Dept. Agric. Exper. Station Bull.*, 19 No. 208.

70. ———. The metabolism and energy transformations healthy man during rest. *Carnegie Instit. Wash. Pu* 1910, No. 126.

71. ———. Respiration calorimeters for studying the resp tory exchange and energy transformations of man. *C negie Instit. Wash. Publ.*, 1910, No. 123.

72. Benedict, F. G., & Cathcart, E. P. Muscular work metabolic study with special reference to the efficiency the human body as a machine. *Carnegie Instit. W Publ.*, 1913, No. 187.

73. Benedict, F. G., & Collins, W. E. A clinical appara for measuring basal metabolism. *Boston Med. & Surg.* 1920, 183, 449—.

74. Benedict, F. G., Emmes, L. E., & Riche, J. A. The in ence of the preceding diet on the respiratory quotient a active digestion has ceased. *Amer. J. Physiol.*, 1911, 383—.

75. Benedict, F. G., & Finn, M. D. Basal metabolism be and after a summer vacation. *Amer. J. Physiol.*, 1928, 665-671.

76. ———. Normal menstruation and gaseous metaboli *Amer. J. Physiol.*, 1928, 86, 59—.

77. BENEDICT, F. G., & JOHNSON, A. Energy loss of young women in light household work. *Proc. Amer. Phil. Soc.*, 1919, 58, 89—.

78. BENEDICT, F. G., MILES, W., ROTH, P., & SMITH, H. M. Human vitality and efficiency under prolonged restricted diet. *Carnegie Instit. Wash. Publ.*, 1919, No. 280.

79. BENEDICT, F. G., & MURSCHHAUSER, H. Energy transformations during horizontal walking. *Carnegie Instit. Wash. Publ.*, 1915, No. 231.

80. BENEDICT, F. G., & PARMENTER, H. S. The energy metabolism of women ascending or descending stairs. *Amer. J. Physiol.*, 1928, 84, 675—.

81. BENEDICT, F. G., & SMITH, H. M. The metabolism of athletes. *J. Biol. Chem.*, 1915, 20, 243—.

82. BENEDICT, F. G., & TOMPKINS, E. H. Respiratory exchange with a description of a respiration apparatus for clinical use. *Boston Med. & Surg. J.*, 1916, 174, 857—, 898—, 939—.

83. Berlin letter. The amount of energy expended by mental work. *J. Amer. Med. Asso.*, 1925, 84, 999—.

84. BIEDL, A. The significance of the internal secretions in disturbances of metabolism and digestion. *Endocrinol.*, 1921, 5, 523-536.

85. BIRCHER, M. E. Das Schwitzen im Lichte des Energieverbrauchs. *Schweiz. med. Woch.*, 1922, 52, 1265—.

86. BLOMBERG, H., & JOHNSON, G. Der Stoffwechsel beim Ping-Pongspielen. *Skand. Arch. f. Physiol.*, 1926, 48, 231-233.

87. BLUNT, K., & DYE, M. Basal metabolism of normal women. *J. Biol. Chem.*, 1921, 47, 69-87.

88. BOCK, A. V., DILL, D. B., HURXTHAL, L. M., LAWRENCE, J. S., COOLIDGE, T. C., DAILEY, M. E., & HENDERSON, L. J. Blood as a physico-chemical system: V. The composition and respiratory exchange of normal human blood during work. *J. Biol. Chem.*, 1927, 73, 749—.

89. BOCK, A. V., DILL, D. B., & TALBOTT, J. H. Studies in muscular activity: I. Determination of the rate of circulation of blood in man at work. *J. Physiol.*, 1928, 66, 121-132.

90. BOCK, A. V., VANCAULAERT, C., DILL, D. B., FÖLLING, A., & HURXTHAL, L. M. Studies in muscular activity: III.

Dynamical changes occurring in man at work. *J. Physiol.*, 1928, 66, 136-161.

91. ———. Studies in muscular activity: IV. The 'steady state' and the respiratory quotient during work. *J. Physiol.*, 1928, 66, 162-174.

92. Bodansky, M. Introduction to physiological chemistry. New York: Wiley, 1927.

93. Boigey, M. Appareil pour le dosage de l'oxygène et de l'acide carbonique de l'air expiré en vue de la détermination du métabolisme. *Presse méd.*, 1925, 33, Pt. 1, 413—.

94. ———. Dosage de l'exercise physique par la mesure de l'acide carbonique émis. *Presse méd.*, 1922, 30, 1189-1193.

95. ———. Évaluation de la dépense physiologique par le dosage de l'acide carbonique émis. Métabolisme pendant l'exercise. *Soc. de méd. milit. fr.*, 1922, 16, 19-22.

96. ———. Influence de l'exercise sur les fonctions circulatoire et respiratoire. *J. méd. fr.*, 1927, 16, 298-305.

97. ———. Influence de l'exercise sur le métabolisme; méthode générale de représentation graphique des echanges respiratoires. *Presse méd.*, 1926, 34, Pt. 2, 724—.

98. ———. Métabolisme et courbe de la dépense physiologique pendant l'exercise. *Mém. et bull. de Soc. méd.-chir. des hôp. et hospices de Bordeaux*, 1927, 2, 83-88.

99. du Bois-Reymond, R. Methoden zum Studium der Atemmechanik (mit Ausschluss der Spirometrie, Pneumatometrie und der Bestimmung des Pleuradrucks). *Handb. d. biol. Arbeitsmeth.*, 1927, Abt. 5, Teil 4, Pt. 2, 1663-1692.

99*a*. du Bois-Reymond, R., & Peltret,—. Der Gaswechsel bei Turnkunststücken. *Arch. f. Physiol.*, 1914, 251-272.

100. Boothby, W. M. Determination of the circulation rate in man at rest and at work. *Amer. J. Physiol.*, 1915, 37, 383—.

101. Boothby, W. M., & Rowntree, L. G. Drugs and basal metabolism. *J. Pharm. & Exper. Therap.*, 1923, 22, 99-108.

102. Boothby, W. M., & Sandiford, I. Basal metabolism. *Physiol. Rev.*, 1924, 4, 69-161.

103. ———. A comparison of the DuBois and the Harris and Benedict standards for the estimation of the basal metabolic rate. *J. Biol. Chem.*, 1922, 54, 767—.

104. ———. Laboratory manual of the technic of basal metabolic rate determinations. Philadelphia: Saunders, 1920. Pp. 117.

105. ———. Summary of the basal metabolism data on 8614 subjects with special reference to the normal standards for the estimation of the basal metabolic rate. *J. Biol. Chem.*, 1922, **54**, 783—.

106. BOOTHBY, W. M., & SANDIFORD, R. B. Nomographic charts for the calculation of the metabolic rate by the gasometer method. *Boston Med. & Surg. J.*, 1921, **185**, 337-354.

107. BORDAS, F., & COURTIER, J. L'hygiène du travail. La prophylaxie de la fatigue et ses avantages sociaux. *Ann. d'hygiène pub., indus., et soc.*, 1927, 5, 327—.

198. BOUNHIOL, J. P. Un sphyximètre. *Bull. Soc. chim. biol.*, 1925, 7, 1139-1143.

109. BOVARD, J. F., & COZENS, F. W. Tests and measurements in physical education 1861-1925. A treatment of the original sources with critical comment. *Univ. Ore. Publ., Phys. Educ. Ser.*, 1926, 1. Pp. 94.

110. BOWEN, B. P., & CARMER, M. E. The effect of a standard exercise upon the oxygen consumption of normal, overweight and chronically ill individuals. *J. Clin. Invest.*, 1925-1926, 2, 299—.

111. BRADFIELD, H. S. The determination of the surface area of women and its use in expressing basal metabolic rate. *Amer. J. Physiol.*, 1927, **82**, 570-576.

112. BRIGGS, H. Physical exertion, fitness and breathing. *J. Physiol.*, 1923, **54**, 292—.

113. ———. Physical work and the human machine. *Eng. & Indus. Manag.*, 1921, 6 (new ser.), 323—.

114. Brown Instrument Co. Brown electric CO_2 meters. *Catalog No. 33 of Brown Instrument Co.*, Philadelphia, Pa.

115. BRUNTON, T. L. Collected papers on circulation and respiration. (2 vols.) 1916.

116. BURGER, G. C. E., & DUSSER DE BARENNE, J. G. Ein Apparat zur graphischen Bestimmung des Gesamtgaswechsels beim Menschen. *Zsch. f. d. ges. exper. Med.*, 1926, 49, 130-146.

117. ———. Vergilijkend onderzoek over de gaswisseling tijdens phasichen en statischen arbeid van den mensch. *Nederl. tijdschr. v. geneesk.*, 1928, 72, IA, 419-472.

118. Caccuri, S. Metabolismo basale e lavoro. *Arch. di pat. e clin. med.*, 1927, 6, 305-327.

119. Cambridge Instrument Co. Cambridge physiological instruments booklet No. V. *Catalog of Cambridge Instrument Co.*, London & New York.

120. Campbell, J. A. Ultra-violet radiation and metabolism with a new method for estimating metabolism. *Proc. Roy. Soc. London*, 1926-27, 99B, 451-461.

121. Campbell, J. M. H., Douglas, G., & Hobson, F. G. The respiratory exchange of man during and after muscular exercise. *Trans. Roy. Soc.*, 1920, B210, 1—.

122. Carpenter, T. M. An apparatus for the exact analysis of air in metabolism investigations with respiratory exchange chambers. *J. Metab. Res.*, 1924, 4, 1-25.

123. ———. A comparison of methods for determining the respiratory exchange of man. *Carnegie Instit. Wash. Publ.*, 1915, No. 216.

124. ———. The increase in metabolism due to the work of typewriting. *J. Biol. Chem.*, 1911, 9, 231—.

125. ———. Tables, factors and formulas for computing respiratory exchange and biological transformations of energy. (2nd ed.) *Carnegie Instit. Wash. Publ.*, 1924, No. 303A.

126. Carpenter, T. M., & Benedict, F. G. The metabolism of man during the work of typewriting. *J. Biol. Chem.*, 1909, 6, 271—.

127. Carpenter, T. M., & Fox, E. L. Absence of stratificaton and rapidity of mixing of carbon dioxide in air samples. *J. Biol. Chem.*, 1927, 73, 379—.

128. Cassinis, V. Rapporto fra ritmo cardiaco e respiratorio durante e dopo il lavoro muscolare e nel riposo. *Giorn. di med. milit.*, 1926, 74, 157-179.

129. ———. Lo scambio gassoso respiratorio dopo ingestione di glicosio nel riposo e nel lavoro muscolare. *Arch. di fisiol.*, 1927, 25, 590—.

130. Cathcart, E. P. The influence of muscle work on protein metabolism. *Physiol. Rev.*, 1925, 5, 225—.

131. ———. Method of estimating energy expenditure by indirect calorimetry. *Roy. Army Med. Corps J.*, 1918, 31, 339—.

132. Cathcart, E. P., Bedale, E. M., Macleod, K., Weather-

HEAD, E., & OVERTON, S. G. The physique of women in industry: a contribution towards the determination of the optimum load. *Rep. Indus. Fatigue Res. Board,* 1927, No. 44. Pp. 137.

133. CATHCART, E. P., & BURNETT, W. A. The influence of muscle work on metabolism in varying conditions of diet. *Proc. Roy. Soc. London,* 1926, B99, 405—.

134. CATHCART, E. P., LOTHIAN, N. V., & GREENWOOD, M. A note on the rate of marching and the expenditure of energy in man. *Roy. Army Med. Corps J.,* 1920, 34, 297—.

135. CATHCART, E. P., & MARKOWITZ, J. The influence of various sugars on the respiratory quotient: a contribution to the significance of the R. Q. *J. Physiol.,* 1927, 63, 309—.

136. CATHCART, E. P., & ORR, J. B. Energy expenditure of the infantry recruit in training. London: H. M. Stationery Office. Pp. 69.

137. CATHCART, E. P., & STEVENSON, A. G. Energy expenditure in man; the cost of positive, negative and static work. *Roy. Army Med. Corps J.,* 1922, 38, 1-8.

138. CATHCART, E. P., & TRAFFORD, F. J. Energy expenditure in minor duties. *J. Physiol.,* 1919, 53, xcix—.

139. CHLOPIN, G. W., JAKOWENKO, W., & WOLCHINSKY, W. Weitere Untersuchungen über den Einfluss der geistigen Tätigkeit auf den respiratorischen Gaswechsel und auf den Energieumsatz. *Arch. f. Hygiene,* 1927, 98, 158-175.

140. COHEN, S. S. The basal metabolism of man in the tropics. *J. Amer. Med. Asso.,* 1922, 78, 297—.

141. COLLIS, E. L., & GREENWOOD, MAJOR. Health of the industrial worker. London: 1921.

142. CONDERO, N. On the alveolar CO_2 tension following vigorous muscular exercise. *Amer. J. Physiol.,* 1925, 77, 91—.

143. CONYBEARE, E. J., & PEMBREY, M. S. The significance of the respiratory quotient. *Guy's Hosp. Rep.,* 1926, 76, 342-350.

144. COOK, M. B., & PEMBREY, M. S. Observations on the effect of muscular exercise upon man. *J. Physiol.,* 1912-1913, 45, 429—.

145. CORLETTE, C. E. On some relations of climate, weather and fat covering to metabolism. *Med. J. Australia,* 1923, 10, Pt. 1, 172-182.

146. Coro, A. J. Contribución al estudio del metabolismo basal en los países tropicales. *Arch. de Soc. estud. clin.*, 1927, 27, 7-20.

147. Cowgill, G. R. An improved procedure for metabolism experiments. *J. Biol. Chem.*, 1923, 56, 725-736.

148. Crawley, S. L. An experimental investigation of recovery from work. *Arch. Psychol.*, 1926, 13, No. 85. Pp. 66.

149. Crofts, E. Effect of ultra-violet radiation on basal metabolism of human subjects. *Amer. J. Hygiene*, 1928, 8, 1020-1023.

150. Crowden, G. P. The physiological cost of the muscular movements involved in barrow work. *Rep. Indus. Fatigue Res. Board*, 1928, No. 50.

151. Curtis, G. M. Beiträge zur Physiologie der Drüsen von Leon Asher, Nr. 91. Fortgesetzte Untersuchungen über den respiratorischen Stoffwechsel bei Arbeit in seiner Beziehung zu den Drüsen mit innerer Sekretion. *Biochem. Zsch.*, 1926, 167, 321-328.

152. Dana, R. T. The human machine in industry. New York: Codex Book Co., 1927. Pp. 307.

153. Dautrebande, D. Étude critique de l'appareil portatif de Benedict pour la détermination du métabolisme de base. *Arch. int. de méd. expér.*, 1928, 4, 327-334.

154. Dautrebande, D., & Davies, H. W. Variations in respiratory exchange with masks of different types. *Edinburgh Med. J.*, 1922, 29 (3rd ser.), 127-135.

155. ———. Le métabolisme basal; étude critique d'une méthode simplifiée. *Bull. Acad. roy. méd. de Belgique*, 1922, 2 (5th ser.), 147-179.

156. Daynes, H. A. The theory of the katharometer. *Proc. Roy. Soc. London*, 1920, A97, 273—.

157. Day, M. E. Influence of mental activities on vasomotor processes. *J. Comp. Psychol.*, 1923, 3, 333—.

158. Dethloff, H. Apparat zur graphischen Registrierung des Sauerstoffverbrauchs und der Kohlensäure-Produktion (with remarks by Prof. Leschke). *Klin. Woch.*, 1925, 4, 2440-2442.

159. Dhar, N. R. The influence of temperature on metabolism and the problem of acclimatization. *J. Phys. Chem.*, 1926, 30, 480-490.

DILL, D. B., & FÖLLING, A. Studies in muscular activity; nomographic description of expired air. *J. Physiol.*, 1928, 66, 133-135.

DILL, D. B., HURXTHAL, L. M., VAN CAULAERT, C., FÖLLING, A., & BOCK, A. The carbon dioxide equilibrium in alveolar air and arterial blood: II. Resting subject. *J. Biol. Chem.*, 1927, 74, 303—.

DILL, D. B., LAWRENCE, J. S., HURXTHAL, L. M., & BOCK, A. V. The carbon dioxide equilibrium in alveolar air and arterial blood: III. Exercising subjects. *J. Biol. Chem.*, 1927, 74, 313—.

DISERENS, C. M. The influence of music on behavior. Princeton, N. J.: Princeton Univ. Press, 1926. Pp. 210.

DIXON, R. F. Psychology studies noise. *Trained Nurse & Hosp. Rev.*, 1927, 78, 253—.

DODGE, R. Mental work. *Psychol. Rev.*, 1913, 20, 1-42.

DOUGLAS, C. G. Coordination of respiration and circulation with variations in bodily activity. *Lancet*, 1927, 2, 213—, 265—.

————. A method for determining the total respiratory exchange in man. *J. Physiol.*, 1911, 42, xvii—.

DOUGLAS, C. G., & HALDANE, J. S. The capacity of the air passages under varying physiological conditions. *J. Physiol.*, 1912, 45, 235—.

DOUGLAS, C. G., HALDANE, J. S., HENDERSON, Y., & SCHNEIDER, E. C. Physiological observations made on Pike's Peak, Colorado, with special reference to adaptation to low barometric pressures. *Phil. Trans. Roy. Soc.*, 1913, B203, 185—.

DOUGLAS, C. G., & PRIESTLY, J. G. Human physiology. London: Oxford Univ. Press, 1924. Pp. 232.

DREYER, G. The assessment of physical fitness. London: Cassell & Co., 1921.

DUBOIS, E. F. Basal metabolism in health and disease. Philadelphia: Lea & Febiger, 1927. Pp. 414.

————. A chart showing graphically the respiratory quotient and the percentage of calories furnished by protein, fat, and carbohydrate. *Proc. Soc. Exper. Biol. & Med.*, 1923-1924, 21, 62—.

————. The proportions in which protein, fat, and car-

bohydrate are metabolized in disease. In *Lectures on nutrition, Mayo foundation lectures,* 1925. Pp. 77—.

175. DuBois, D., & DuBois, E. F. A formula to estimate the approximate surface area if height and weight be known. *Arch. Internal Med.,* 1916, 17, 863—. Also in *Russell Sage Reps.,* 1916, 1, 11—.

176. Durig, A. Die Ermüdung im praktischen Betrieb. In *Körper und Arbeit,* ed. by E. Atzler. Leipzig: Thieme, 1927. Pp. 488-647.

177. ———. Über den Gaswechsel beim Gehen auf horizontaler Bahn, and Über den Gaswechsel beim Gehen auf ansteigender Bahn. *Denkschr. d. Kaiserl. Akad. d. Wiss.,* 1911, 86, 241—, 293—.

178. ———. Die Theorie der Ermüdung. In *Körper und Arbeit,* ed. by E. Atzler. Leipzig: Thieme, 1927. Pp. 196-328.

179. ———. Über das Verhalten der Atemmechanik und der Alveolartension. *Denskschr. d. Kaiserl. Akad. d. Wiss.,* 1911, 86, 349-481.

180. Dusser de Barenne, J. G., & Burger, G. C. E. Ein Apparat zur graphischen Bestimmung des Sauerstoffverbrauchs und der Kohlensäureabgabe, somit des Gesamtgaswechsels beim Menschen. *Klin. Woch.,* 1925, 4, 68-71.

181. ———. Graphische Methoden zur Bestimmung des Gesamtgaswechsels des Menschen im Arbeitsversuch. *Arch. f. d. ges. Physiol.,* 1927, 218, 222-238.

182. ———. Untersuchungen über den Gaswechsel des Menschen bei statischer Arbeit. *Arch. f. d. ges. Physiol.,* 1927, 218, 239-260.

183. Efimov, V. V. [Energetics of work.] *Central'nyj Instit. Truda v. c. s. P. S.-Cit.,* Petrovka, 1924, No. 10.

184. ———. [Study of fatigue as a method of scientific task setting.] *Central'nyj Instit. Tuda v. c. s. P. S.-Cit.,* Petorvka, 1923, Nos. 8-9.

185. Emmes, L. E., & Riche, J. A. The respiratory exchanges as affected by body position. *Amer. J. Physiol.,* 1910, 27, 406—.

186. Evans, C. L. Recent advances in physiology. (3rd ed.) Philadelphia: Blakiston, 1928. Pp. 403.

187. Farkas, G., & Geldrich, J. Über den Energieverbrauch beim Orgelspiel. *Arch. f. Hygiene,* 1928, 99, 52-60.

188. FARKAS, G., GELDRICH, J., & SZAKÁLL, A. Über den Energieumsatz beim Ernten. *Arbeitsphysiol.*, 1929, 1, 466-479.

189. FARMER, E. Interpretation and plotting of output curves. *Brit. J. Psychol.*, 1923, 13, 308—.

190. FENN, W. O. The gas exchange of isolated muscles during stimulation and recovery. *Amer. J. Physiol.*, 1927, 83, 309—.

191. ———. A sensitive method for measuring carbon dioxide. *Proc. Soc. Exper. Biol. & Med.*, 1926, 23, 714—.

192. FLEMING, G. B. The influence of growth on the basal metabolism of children. *Amer. J. Dis. Child.*, 1923, 25, 85-88.

193. FLICKINGER, R. Zur Frage der Stoffwechseländerung durch Bestrahlung. *Dtsch. med. Woch.*, 1926, 52, 1501—.

194. FLORENCE, P. S. Economics of fatigue and unrest and the efficiency of labor in English and American industry, 1924. Pp. 426.

195. ———. Individual variations in efficiency and the analysis of the work curve. *Arch. f. Weltwirtschaft*, Kiel University, 1924.

196. ———. Use of factory statistics in the investigation of industrial fatigue. *Columbia Univ. Stud. Hist., Econ., & Pub. Law*, 1918, 81, No. 3.

197. FRANK, R., & HERZGER, R. Beitrag zur Untersuchung des Gasstoffwechsels beim Menschen. *Klin. Woch.*, 1926, 5, 834-837.

198. FRIES, J. A. The R. Q. and its uncertainty. *Amer. J. Physiol.*, 1920, 51, 202—.

199. FRUMERIE, K. Ueber das Verhaeltniss des Ermuedungsgefuehls zur CO_2-abgabe bei statischer Muskelarbeit. *Skand. Arch. f. Physiol.*, 1913, 30, 409—.

200. FULTON, W. B. An improved air valve for apparatus used in basal metabolic work. *Arch. Internal Med.*, 1924, 33, 497—.

201. FURUSAWA, K. Muscular activity and carbohydrate metabolism in the normal individual. *Proc. Roy. Soc. London*, 1925, 98B, 65—.

202. ———. Muscular exercise, lactic acid and the supply and utilization of oxygen: XIII. The gaseous exchange of restricted muscular exercise in man. *Proc. Roy. Soc. London*, 1926, 99B, 155-166.

203. ———. A spirometer method of studying continuously the gaseous metabolism of man during and after exercise. *Proc. Roy. Soc. London,* 1926, 99B, 148—.

204. FURUSAWA, K., HILL, A. V., & PARKINSON, J. L. The dynamics of 'sprint' running. *Proc. Roy. Soc. London,* 1927, 102B, 29—.

205. ———. The energy used in 'sprint' running. *Proc. Roy. Soc. London,* 1927, 102B, 43—.

206. GAIRNS, S., & O'BRIEN, M. K. A discussion of the method of A. D. Waller for computing energy output and results of experiments determining energy expenditure during some household tasks. *J. Indus. Hygiene,* 1922, 4, 283—.

207. GAPPELLEN, L., & NOYONS, A. K. Méthode gazo-analytique à voie ouverte servant à la détermination du métabolisme. *C. r. Soc. de biol.,* 1925, 93, 1530-1533.

208. GAUSS, H. Corrections for gas volumes for altitudes 700 to 600 mm. *J. Metab. Res.,* 1924, 4, 415-422.

209. GELDRICH, J. Ueber den Energieverbrauch beim Reiten. *Biochem. Zsch.,* 1927, 188, 1-14.

210. GEPHART, F. C., DUBOIS, E. F., & LUSK, G. Clinical calorimetry. The number of places of significant figures in the data of metabolism experiments. *J. Biol. Chem.,* 1916, 27, 217—.

211. GESELL, R. Another outlook on the chemical regulation of respiration. *Science,* 1926, 63, 58-62.

212. ———. The chemical regulation of respiration. *Physiol. Rev.,* 1925, 5, 551—.

213. GESELL, R., & MCGINTY, D. A. Regulation of respiration, continuous electrometric methods of recording changes in expired carbon dioxide and oxygen. *Amer. J. Physiol.,* 1926, 79, 72—.

214. GIESE, F. Methoden der Wirtschaftspsychologie. Berlin-Wien: Urban und Schwarzenberg, 1927. Pp. 631.

215. GILLESPIE, R. D. The relative influence of mental and muscular work on the pulse rate and blood pressure. *J. Physiol.,* 1924, 58, 425—.

216. GLIBERT, D. J. Influence of industrial noises. *J. Indus. Hygiene,* 1922, 264—.

217. GMEINER, G. CO_2 und O_2 Analysen mit dem Haldane Apparat mittels automatischer Schwenkung. *Biochem. Zsch.,* 1927, 188, 285-290.

218. GOIFFON, R. Appareil simple pour le prélèvement du CO_2 alvéolaire. *C. r. Soc. de biol.*, 1928, **99**, 377—.

219. GOLDBERG, A. F., & LEPSKAIA, M. V. Les altérations des globules blancs au cours du travail physique et intellectuel. *J. de physiol.*, 1926, **24**, 715—.

220. GOLDMARK, J. Fatigue and efficiency. New York: Russell Sage Foundation, 1912.

221. GOLDSTEIN, J. D., & DUBOIS, E. L. The effect on the circulation in man of rebreathing different concentrations of carbon dioxide. *Amer. J. Physiol.*, 1927, **81**, 650-659.

222. GORDON, B. Effect of exercise on the circulation, sugar metabolism and other factors. *Northwest Med.*, 1926, **25**, 376—.

223. GÖTTSCHE, O. Gasstoffwechseluntersuchungen im Kindesalter. Die Pubertätsreaction. *Monatssch. f. Kinderhk.*, 1926, **32**, 22-36.

224. GRAF, O. Die Arbeitspause in Theorie und Praxis. *Psychol. Arbeit*, 1928, **9**, 563-683.

225. GRAFE, E. Die pathologische Physiologie des Gesamtstoff- und Kraftwechsels bei der Ernährung des Menschen. *Erg. d. Physiol.*, 1923, **21**, Pt. 2. Pp. 557. Also Munich: Bergmann, 1923.

226. ————. Ein Respirationsapparat. *Zsch. f. physiol. Chem.*, 1910, **65**, 1—.

227. GRAFE, E., & MAYER, L. Ueber den Einfluss der Affekte auf den Gesamtstoffwechsel. Untersuchungen in der Hypnose. *Zsch. f. d. ges. Neur. u. Psychiat.*, 1923, **86**, 247-253.

228. GRAFE, E., STRIECK, F., & OTTO-MARTIENSEN, J. Ein Universalrespirationsapparat. *Zsch. f. d. ges. exper. Med.*, 1927, **54**, 613-630.

229. GRAFE, E., & TRAUMANN, E. Zur Frage des Einflusses psychischer Depressionen und der Vorstellung schwerer Muskelarbeit auf den Stoffwechsel (Untersuchungen in der Hypnose). *Zsch. f. d. ges. Neur. u. Psychiat.*, 1920, **62**, 237-252.

230. GREENWOOD, M., HODSON, C., & TEBB, A. E. Metabolism of female munition workers. *Brit. Med. J.*, 1919, No. 3048, 682—.

231. ————. Report on the metabolism of female munition workers. *Proc. Roy. Soc. London*, 1919, **91B**, 62—.

232. GREENWOOD, M., & NEWBOLD, E. M. On the estimation of metabolism from determinations of CO_2 production and on the estimation of external work from the respiratory metabolism. *J. Hygiene,* 1923, 21, 440—.

233. GRIFFITH, F. R., PUCHER, G. W., BROWNELL, K. A., CARMER, M. E., & KLEIN, J. D. Seasonal periodicity in man: I. Basal metabolism and respiratory vascular changes *Amer. J. Physiol.,* 1927, 81, 483—.

234. GRIFFITHS, E. A gas analysis instrument based on sound velocity measurement. *Proc. Phys. London,* 1927, 39, 300—.

235. GUILLAUME, A. C. Méthode spectroscopique d'estimation du métabolisme fondamental. *J. méd. fr.,* 1923, 12, 262-263.

236. GULLICHSEN, R., & SOISALON-SOININEN, J. L. Über die Kohlenstoffabgabe des Menschen beim Fechten und Ringen. *Skand. Arch. f. Physiol.,* 1921, 41, 188—.

237. GUSTAFSON, F. L., & BENEDICT, F. G. The seasonal variation in basal metabolism. *Amer. J. Physiol.,* 1928, 86, 43-58.

238. GUTHRIE, C. C. A simple and accurate metabolism spirometer. *Arch. Internal. Med.,* 1921, 28, 841—.

239. HAFKESBRING, R., & COLLETT, M. Day to day variations in basal metabolism of women. *Amer. J. Physiol.,* 1924, 70, 73-85.

240. HAGEDORN, H. C. An apparatus for the graphic recording of oxygen consumption and carbon dioxide output, especially adapted for clinical work. *Biochem. J.,* 1924, 18, 1301—.

241. HAGGARD, H. W. Fatigue from a new point of view: physiological maintenance of the human machine in industry. *Mfg. Indus.,* 1926, 11, 363—.

242. HALDANE, J. S. Methods of air analysis. (2nd ed.) 1918.

243. ————. Respiration. New Haven, Conn.: Yale Univ. Press, 1922. Pp. 427.

244. HAMMARSTEN, E. Untersuchungen über die Kohlensäureabgabe bei statischer und negativer Muskelarbeit. *Skand. Arch. f. Physiol.,* 1912, 26, 212-220.

245. HANNAM, G. C. Reduction of office noise. *Amer. Manag. Asso., Office Exec. Ser.,* 1926, No. 18.

246. HANNAN, R. R., & LYMAN, R. S. A graphic method for re-

cording oxygen consumption. *Johns Hopkins Hosp. Bull.*, 1923, 34, 241—.

HANRIOT, M. M., & RICHET, C. Des échanges respiratoires chez l'homme. *Ann. de chim. et de phys.*, 1891, 22, 495—.

HANSEN, E. Der Einfluss des Arbeitstempos auf Respiration und Kreislauf. *Skand. Arch. f. Physiol.*, 1928, 54, 50-78.

HANSEN, E., HVORSLEV, C. M., & LINDHARD, J. 'Aktive' und 'passive' Muskelkraft. *Skand. Arch. f. Physiol.*, 1928, 54, 99-115.

HARDIKAR, S. W. The action of quinine on protein metabolism, respiratory exchange, and heart function. *Brit. Med. J.*, 1924, 1, Pt. 2, 1125-1127. Also in *J. Pharm. & Exper. Therap.*, 1925-1926, 25, 175-218.

HARRIS, J. A., & BENEDICT, F. G. A biometric study of basal metabolism in man. *Carnegie Instit. Wash. Publ.*, 1919, No. 279.

————. The variation and the statistical constants of basal metabolism in men. *J. Biol. Chem.*, 1921, 46, 257-279.

HARTWELL, G., & TWEEDY, N. Some effects of muscular exercise on women. *J. Physiol.*, 1913, 46, ix-xi.

HASTINGS, A. B. An investigation of changes in the blood resulting from fatigue. *U. S. Treas. Dept., Pub. Health Rep.*, No. 546, 34, No. 31, 1682-1691.

————. The physiology of fatigue. *U. S. Treas. Dept., Pub. Health Bull.*, 1921, No. 117.

HAWK, P. B., & BERGHEIM, O. Practical physiological chemistry. (9th ed.) Philadelphia: Blakiston, 1926.

HECHT, V. Ueber den Einfluss mittlerer Höhenlagen auf Kreislaufs- und Atmungsorgane bei raschem Höhenwechsel. *Wien. med. Woch.*, 1927, 77, 1048-1053.

HEDON, L. Quelques considérations sur le métabolisme basal. Calcul de la dépense d'énergie par unité de poids et par unité de surface. *Bull. Soc. de scient. méd. et biol. de Montpellier*, 1922-1923, 4, 300-304.

HEINBECKER, P. Studies on the metabolism of Eskimos. *J. Biol. Chem.*, 1928, 80, 461-475.

HELMREICH, E., & WAGNER, R. Die graphische Bestimmung der Kohlensäureausscheidung und des respiratorischen Quotienten mit dem Differenzspirometer. *Biochem. Zsch.*, 1924, 149, 560-565.

261. ———. Über ein neues Prinzip zur indirekten Bestimmung der Kohlensäure und des respiratorischen Quotienten mit dem Differenzspirometer. *Klin. Woch.*, 1924, 3, 406-407.

262. HENDERSON, L. J., DILL, D. B., VAN CAULAERT, C., FÖLLING, A., & COOLIDGE, T. C. The steady state of heavy work. *J. Biol. Chem.*, 1927, 74, xxxvi-xxxvii.

263. HENDERSON, Y. The Haldane gas analyzer. *J. Biol. Chem.*, 1918, 33, 31—.

264. ———. The respiratory exchange and indirect calorimetry. *J. Biol. Chem.*, 1918, 33, 47—.

265. HENDERSON, Y., & HAGGARD, H. W. The maximum of human power and its fuel. *Amer. J. Physiol.*, 1925, 72, 264-282.

266. HENDERSON, Y., & HAGGARD, H. W. Noxious gases and the principles of respiration influencing their action. (*Monog. Ser.*, No. 35.) New York: Chemical Catalog Co. Pp. 220.

267. HENDERSON, Y., & PRINCE, A. L. The oxygen pulse and the systolic discharge. *Amer. J. Physiol.*, 1914, 35, 106—.

268. HENDRY, M. F., CARPENTER, T. M., & EMMES, L. E. Gaseous exchange with unpracticed subjects and two respiration apparatus employing three breathing appliances. *Boston Med. & Surg. J.*, 1919, 181, 284-296, 334-344, 368-371.

269. HERBST, R. Sport und Arbeit. In *Körper und Arbeit*, ed. by E. Atzler. Leipzig: Thieme, 1927. Pp. 716-733.

270. HERBST, R., & NEBULONI, A. Über den Einfluss der Ermüdung auf Gaswechsel, Puls und Atmung. *Zsch. f. d. ges. exper. Med.*, 1927, 57, 450-469.

271. HERXHEIMER, H. Ein einfacher Apparat zur Bestimmung des Grundumsatzes. *Dtsch. med. Woch.*, 1925, 51, 1143—.

272. HERXHEIMER, H., & KOST, R. Das Verhältnis von Sauerstoffaufnahme und Kohlensäureausscheidung zur Ventilation bei harter Muskelarbeit. *Zsch. f. klin. Med.*, 1928, 108, 240-247.

273. HERXHEIMER, H., WISSING, E., & WOLFF, E. Spätwirkungen erschöpfender Muskelarbeit auf den Sauerstoffverbrauch; Versuche im Höhenklima. *Zsch. f. d. ges. exper. Med.*, 1926, 52, 447-463.

274. Hewitt, E. M., & Bedale, E. M. A study of the comparative physiological costs of different methods of weight carrying by women. *Ann. Rep., Chief Inspector Factories,* 1923, 106—.

275. Hichens, W. L., & Vernon, H. M. Discussion on the physical and mental effects of fatigue in modern industry. *Brit. Med. J.,* 1926, ii, 479-481.

276. Hietanen, A., Nikkinen, M., Nygssölä, H., & Sternberg, G. Der Stoffwechsel beim Gehen auf glatter Fläche. *Skand. Arch. f. Physiol.,* 1928, 54, 145-148.

277. Higgins, H. L. The influence of food, posture and other factors on the alveolar carbon dioxide tension in man. *Amer. J. Physiol.,* 1913, 34, 114—.

278. Higley, G. O., & Bowen, W. P. Changes in the excretion of carbon dioxide resulting from bicycling. *Amer. J. Physiol.,* 1905, 12, 311—.

279. Hill, A. V. An electrical method (katharometer) for measuring CO_2 in respired gases. *J. Physiol.,* 1922, 56, xx—.

280. ———. Living machinery. New York: Harcourt, Brace, 1927.

281. ———. Muscular activity. Baltimore, Md.: Williams & Wilkins, 1926.

282. ———. Muscular activity and carbohydrate metabolism. *Science,* 1924, N. S. 60, 505-515.

283. ———. Muscular activity and carbohydrate metabolism. In *Lectures on nutrition, Mayo foundation lectures,* 1925. Pp. 109—.

284. ———. Muscular movement in man. New York: McGraw Hill, 1927.

285. Hill, A. V., Long, C. N. H., & Lupton, H. Muscular exercise, lactic acid and the supply and utilization of oxygen. *Proc. Roy. Soc. London,* 1924, 96B, 438—; 97B, 84—, 155—.

286. Hill, A. V., & Lupton, H. Muscular exercise, lactic acid and utilization of oxygen. *Quar. J. Med.,* 1923, 16, 135—.

287. ———. The oxygen consumption during running. *J. Physiol.,* 1922, 56, xxxii—.

288. Hill, L. The science of ventilation and open air treatment. *Med. Res. Coun., Spec. Rep.,* London, 1919, Ser. No. 32, Pt. 1.

289. HILL, L., & CAMPBELL, J. A. The effect of atmospheric cooling power on the pulse-rate and on the efficiency during muscular exercise. *J. Physiol.*, 1922, 56, xlix—.

290. ———. Observations on the resting metabolism of children and adults in Switzerland. *Brit. Med. J.*, 1922, i, 385—.

291. ———. The physiological cost of muscular work. *Brit. Med. J.*, 1921, i, 733—.

292. HIMWICH, H. E., & CASTLE, W. B. Studies in the metabolism of muscle: I. The respiratory quotient of resting muscle. *Amer. J. Physiol.*, 1927, 83, 92—.

293. HIMWICH, H. E., & ROSE, M. I. The respiratory quotient of exercising muscle. *Proc. Soc. Exper. Biol. & Med.*, 1926, 24, 169—.

294. HINDMARSH, E. M. Basal metabolic rate of students in Sydney, N. S. W., with a discussion on methods of determining basal metabolism. *Australian J. Exper. Biol. & Med. Sci.*, 1927, 4, 225-268.

295. HOLLINGSWORTH, M. W. A simple method for computation of metabolic determinations by the use of the slide rule. *J. Lab. & Clin. Med.*, 1922-1923, 8, 538-541.

296. HOOGHEIMSTRA, J. S. The importance of silence. In *Industrial leadership*. Soc. Indus. Eng., 1921.

297. HOPKINS, F. G. The chemical dynamics of muscle. *Johns Hopkins Hosp. Bull.*, 1921, 32.

298. HOUGH, T. The influence of muscular activity upon the alveolar tensions of oxygen and of CO_2. *Amer. J. Physiol.*, 1912, 30, 18—.

299. HOWELL, W. H. A textbook of physiology. (10th ed.) Philadelphia: Saunders, 1927.

300. ILZHÖFER, H. Über den Einfluss der geistigen Arbeit auf den Energieverbrauch. *Arch. f. Hygiene*, 1924, 94, 317-328.

301. ———. Über den Energieverbrauch beim Maschinenschreiben. *Arch. f. Hygiene*, 1925, 95, 245-262.

302. International Labor Office. The protection of eye-sight in industry. *Int. Labor Office Stud. & Rep.*, 1923, Ser. F., No. 6.

303. JACKSON, J. B. Relation of pulse rate and pulse pressure to basal metabolism. *J. Mich. State Med. Soc.*, 1926, 25, 565-567.

304. JAHN, D. Die spezifisch-dynamische Wirkung der Nahrung und die Gesetze des Gasstoffwechsels. *Dtsch. Arch. f. klin. Med.*, 1928, 159, 335-361.

305. JANET, H. Principe, graphiques et tables de corrections pour le calcul de résultat dans une épreuve clinique de métabolisme basal. *J. de physiol.*, 1922, 20, 366-375.

306. JAQUET, A. Der respiratorische Gaswechsel. *Erg. d. Physiol.*, 1903, 2, Pt. 1, 457-573.

307. JASCHWILI, D. Über den Einfluss der Trägheit der bewegten Massen auf die Ökonomie der Kurbelarbeit. *Arbeitsphysiol.*, 1928, 1, 198-202.

308. JOHANSSON, J. E. Methodik des Energiestoffwechsels. *Abderhalden's Handb. d. biochem. Arbeitsmeth.*, 1930, 3, Teil 2, 1114-1170.

309. ———. Methodik des Energiestoffwechsels. *Handb. d. biol. Arbeitsmeth.*, 1925, Abt. 4., Teil 9, 331-362.

310. ———. Untersuchungen über die Kohlensäuereabgabe bei Muskelthätigkeit. *Skand. Arch. f. Physiol.*, 1901, 11, 273-307.

311. JOHANSSON, J. E., & KORAEN, G. Untersuchungen über die Kohlensäureabgabe bei statischer und negativer Muskelthätigkeit. *Skand. Arch. f. Physiol.*, 1902, 13, 229-250.

312. JONES, H. M. Basal metabolism determination and its technical difficulties. *J. Lab. & Clin. Med.*, 1921-1922, 7, 191-198.

313. ———. A simple device for measuring basal metabolism. *J. Amer. Med. Asso.*, 1920, 75, 538—. Also in *Arch. Internal Med.*, 1921, 27, 48—.

314. KAHN, H. The vocabulary of metabolism. *Science*, 1922, N. S. 55, 704—.

315. KAUP, J., & GROSSE, A. Energetisches Oberflächengesetz oder ein neues Funktionsgesetz der inneren Organisaton? *Klin. Woch.*, 1927, 6, 2184—, 2223—.

316. ———. Energieaufwand, Herzleistung und Erholungsquotient im Training. *Münch. med. Woch.*, 1927, 74, 1353-1357.

317. ———. Kreislauf, Sauerstoffausnützung und Erholungsquotient menschlicher Arbeit nach teilweise neuer Methodik. *Münch. med. Woch.*, 1926, 73, 1873-1875, 1938-1941.

318. Kestner, O., & Knipping, H. W. Die Ernährung bei geistiger Arbeit. *Klin. Woch.*, 1922, 1, 1353-1354.

319. Kestner, O., & Schadow, H. Strahlung, Atmung und Gaswechsel. Versuche am Jungfraujoch. *Arch. f. d. ges. Physiol.*, 1927, 217, 491-508.

320. Kharina-Marinucci, R. Sull' indice di sufficienza respiatoria volumetrica. *Boll. d. Soc. ital. d. biol. sper.*, 1926, 1, 184-186.

321. King, J. T. Basal metabolism. Baltimore, Md.: Williams & Wilkins, 1924.

322. ———. Determination of the basal metabolism from the carbon dioxide elimination. *Johns Hopkins Hosp. Bull.*, 1921, 32, 277—.

323. King, J. T., & Cross, V. B. Superventilation and carbon dioxide elimination. *Johns Hopkins Hosp. Bull.*, 1923, 34, 349—.

324. Kleiber, M., & Wirth, A. Ein Gasanalysenautomat für Respirationsversuche. *Biochem. Zsch.*, 1928, 192, 241—.

325. Klein, W., & Steuber, M. Die gasanalytische Methodik des dynamischen Stoffwechsels. Leipzig: Thieme, 1925. Pp. 99.

326. Klingendahl, P. O., & Pesonen, N. Die Einwirkung der Tragweise auf den Stoffwechsel. *Skand. Arch. f. Physiol.*, 1928, 54, 169-174.

327. Knipping, H. W. Beitrag zur gasanalytischen Technik in der Medizin. *Zsch. f. d. ges. exper. Med.*, 1926, 53, 1-16.

328. ———. Beitrag zur Technik der Gasstoffwechseluntersuchung. *Zsch. f. d. ges. exper. Med.*, 1925, 47, 1-3.

329. ———. Beitrag zur Technik der Gasstoffwechseluntersuchung. *Zsch. f. physiol. Chem.*, 1925, 145, 154-176.

330. ———. Beitrag zur Technik der Gasstoffwechseluntersuchung. *Münch. med. Woch.*, 1924, 71, 1539—. Also in *Zsch. f. d. ges. exper. Med.*, 1924, 43, 656-658.

331. ———. Beitrag zur Technik der Gasstoffwechseluntersuchung in der Klinik. *Dtsch. Arch. f. klin. Med.*, 1924, 145, 179-195.

332. ———. Ein einfacher Apparat zur exakten Gasstoffwechselbestimmung. *Zsch. f. d. ges. exper. Med.*, 1924, 40, 363-373.

333. ———. Ein einfacher Apparat zur exakten Gasstoffwech-

seluntersuchung in der Klinik und ärztlichen Praxis. *Münch. med. Woch.*, 1924, 71, 553—.

334. ———. Ergebnisse der Gasstoffwechseluntersuchung für die Klinik. *Klin. Woch.*, 1928, 7, 49—.

335. ———. Respiratorischer Gaswechsel, Blutreaktion und Blutphosphorsäurespiegel bei geistiger Arbeit. *Zsch. f. Biol.*, 1923, 77, 165-174.

336. ———. Über die Bestimmung der Kohlensäurespannung in der Alveolarluft. *Zsch. f. physiol. Chem.*, 1924, 141, 1-10.

337. ———. Ueber die Kontrolle der Gasstoffwechseluntersuchung. *Zsch. f. d. ges. exper. Med.*, 1926, 50, 345-358.

338. ———. Ueber die kurzfristigen Gasstoffwechseluntersuchungen. *Klin Woch.*, 1926, 5, 677—.

339. ———. Ueber die Vereinfachung der klinischen Gasstoffwechseluntersuchung. *Münch. med. Woch.*, 1924, 71, Pt. 2, 1168—.

340. ———. Zur Technik der klinischen Gasstoffwechseluntersuchung. *Münch. med. Woch.*, 1925, 72, 2095-2097.

341. ———. Zur Technik der langdauernden experimentellen und der klinischen Gasstoffwechseluntersuchungen. *Zsch. f. d. ges. exper. Med.*, 1927, 57, 433—.

342. KNIPPING, H. W., & KOWITZ, H. L. Klinische Gasstoffwechseltechnik. Berlin: Springer, 1928. Pp. 193.

343. KNIPPING, H. W., & RONA, P. Praktikum der physiologischen Chemie: III. Stoffwechsel und Energiewechsel. Berlin: Springer, 1928. Pp. 268.

344. KNOLL, W. Über äussere und innere Atmung beim Sport. *Schweiz. med. Woch.*, 1927, 157, 361-368.

345. KOMMERELL, B. Die Douglassche Sackmethode zur Bestimmung des Stoffumsatzes. Vereinfachte Versuchsberechnung mit Hilfe nomographischer Tafeln. *Zsch. f. d. ges. exper. Med.*, 1927, 56, 740-747.

346. KORNHAUSER, A. W. The effect of noise on office output. *Indus. Psychol.*, 1927, 2, 621—.

347. KOST, R. Beiträge zur Stoffwechselwirkung der Massage; der Einfluss der Massage auf den Sauerstoffverbrauch des ausgeruhten Organismus. *Zsch. f. d. ges. phys. Therap.*, 1927, 33, 1-11.

348. KRAUSS, E. Lehrbuch der Stoffwechselmethodik. Teil I.

Methodik des Energie- und Stoffwechsels. Leipzig: Hirzel, 1928. Pp. 388—.

349. KROGH, A. A bicycle ergometer and respiration apparatus for the experimental study of muscular work. *Skand. Arch. f. Physiol.*, 1913, 30, 375—.

350. ———. Determination of standard (basal) metabolism of patients by a recording apparatus. *Boston Med. & Surg. J.*, 1923, 189, 313-317.

351. ———. Energy exchange in man. *J. Physiol.*, 1919, 52, lxxiv—.

352. ———. Et Respirationsapparat til klinisk Bestmelse af menneskets Energiomsaetning. *Ugeskr. f. Laeger*, 1922, 84, Pt. I, 525—.

353. ———. Ein Respirationsapparat zur klinischen Bestimmung des Energiumsatzes des Menschen. *Wien. klin. Woch.*, 1922, 35, 290—.

354. ———. Respiratory exchanges of animals and man. New York: Longmans, Green, 1916. Pp. 173.

355. ———. Sur un appareil respiratoire enregistreur, servant à déterminer l'absorption d'oxygène et les échanges caloriques chez l'homme. *C. r. Soc. de biol.*, 1922, 87, 485-461.

356. KROGH, A., & LINDHARD, J. The regulation of respiration and circulation during the initial stages of muscular work. *J. Physiol.*, 1913, 47, 112—.

357. ———. Relative value of fat and carbohydrate as sources of muscular energy. *Biochem. J.*, 1920, 14, 290—.

358. KUNDE, M. M. After effects of prolonged fasting. *J. Metab. Res.*, 1923, 3, 399-449.

359. LABBÉ, M., & STEVENIN, H. Technique de la mesure des échanges respiratoires. *Presse méd.*, 1922, 30, 841-843.

360. LAIRD, D. A. The measurement of the effects of noise upon working efficiency. *J. Indus. Hygiene*, 1927, 9, 431—.

360*a*. ———. Experiments on the physiological cost of noise. *J. Nat. Instit. Indus. Psychol.*, 1929, 4, 251-258.

361. LAIRD, D. A., & WHEELER, W. H., JR. What it costs to lose sleep. *Indus. Psychol.*, 1926, 1, 694—.

362. LANDIS, C. Studies on emotional reactions: IV. Metabolic rate. *Amer. J. Physiol.*, 1925, 74, 188—.

363. LANGWORTHY, C. F., & BAROTT, H. G. Energy expenditure in household tasks. *Amer. J. Physiol.*, 1920, 52, 400—.

364. ———. Energy expenditure in sewing. *Science*, 1921, 54, 446—.

365. ———. Energy expenditure in sewing. *Amer. J. Physiol.*, 1922, 59, 376—.

366. LANGWORTHY, C. F., & MILNER, R. D. An improved respiration calorimeter for use in experiments with man. *J. Agric. Res.*, 1915, 5, 299—.

367. LAQUER, A., & GOTTHEIL, C. Über das Verhalten des Gasstoffwechsels im Kohlensäurebade und im einfachen indifferenten Bade. *Zsch. f. d. ges. phys. Therap.*, 1927, 33, 207-213.

368. LAURENS, H. The physiological effects of radiation. *Physiol. Rev.*, 1928, 8, 1-91.

369. LEDENT, R. Les effets circulatoires de l'exercise (leçon de physiologie médico-sportive). *J. de méd. de Paris*, 1926, 45, 689-691.

370. LEDIG, P. G., & LYMAN, R. S. Adaptation of thermal conductivity method to analysis of respiratory gases. *J. Clin. Invest.*, 1927, 4, 495—.

371. LEE, F. S., & VANBUSKIRK, J. D. An examination of certain proposed tests for fatigue. *Amer. J. Physiol.*, 1923, 63, 185—.

372. LEEDS & NORTHRUP. CO_2 meters (electrical). *Bull.* No. 781, *Power Plant Ser., Leeds & Northrup, Philadelphia*, 1927, No. 5.

373. LEFEVRE, J. Chaleur animale et bioénergétique. Paris: Masson, 1911. Pp. 1107.

374. LEHMANN, G. Arbeitsphysiologische Studien IV. *Arch. f. d. ges. Physiol.*, 1927, 215, 329-364.

375. LEHMANN, G. & MÜLLER, E. A. Ein praktischer Respirationsapparat für die Klinik. *Münch. med. Woch.*, 1927, 74, 1714-1715.

376. LIEBENOW, R. Spätwirkungen erschöpfender Muskelarbeit auf den Sauerstoffverbrauch: IV. Der Verlauf der Erholungskurve unmittelbar nach der Anstrengung *Zsch. f. d. ges. exper. Med.*, 1928, 59, 49-60.

377. VON LIEBERMANN, L., SR. Energiebedarf und mechanisches Äquivalent der geistigen Arbeit. *Biochem. Zsch.*, 1926, 173, 181-189.

378. LIEBESNY, P., & SCHWARZ, H. Registriermethode zur gleichzeitigen Bestimmung des Sauerstoffverbrauchs und der Kohlensäureabgabe. *Biochem. Zsch.*, 1924, 154, 191-194.

379. LILJESTRAND, G., & WOLLIN, G. Über den Einfluss der Körperstellung auf die Atmung des Menschen. *Skand. Arch. f. Physiol.*, 1913, 30, 199-229.

380. LINDHARD, J. Circulation after cessation of work; calculation of circulation rate experiments according to nitrous oxide method. *J. Physiol.*, 1922, 57, 17—.

381. ———. The seasonal periodicity in respiration. *Skand. Arch. f. Physiol.*, 1912, 26, 221-315.

382. ———. Über den respiratorischen Quotienten bei Muskelarbeit von kurzer Dauer. *Skand. Arch. f. Physiol.*, 1928, 54, 79-98.

383. LOEWY, A. Die Gase des Körpers und der Gaswechsel: II. Der respiratorische und der Gesamtumsatz. *Oppenheimer's Handb. d. Biochem.*, 1911, 4, Pt. 1, 133-284.

384. LOEWY, A., & KNOLL, W. Weitere Untersuchungen über den Energieverbrauch beim Skilauf. Zugleich ein Beitrag zur Wirkung des Trainings. *Zsch. f. Hygiene u. Infectionskrankheiten*, 1925, 104, 738-751.

385. LONG, C. N. H. Muscular exercise, lactic acid and the supply and utilization of oxygen: XIV. The relation in man between the oxygen intake during exercise and the lactic acid content of the muscles. *Proc. Roy. Soc. London*, 1926, 99B, 167-172.

386. LUNDSGAARD, C., & MÖLLER, E. Investigations on the immediate effect of heavy exercise on some phases of circulation and respiration in normal individuals. *J. Biol. Chem.*, 1923, 55, 315—, 477—, 599—.

387. LUSK, G. The elements of the science of nutrition. (3rd ed.) Philadelphia: Saunders, 1917.

388. ———. Die Energiequelle bei der Muskelarbeit. *Biochem. Zsch.*, 1925, 156, 334-342.

389. ———. The fundamental basis of nutrition. New Haven, Conn.: Yale Univ. Press, 1923.

390. ———. Fundamental ideas regarding basal metabolism. *J. Amer. Med. Asso.*, 1921, 77, 250-252.

391. ———. A history of metabolism. *Endocrinol. & Metab.*, 1922, 3, 3—.

392. ———. Principes fondamentaux concernant le métabolisme basal. *Gaz. des hôp.*, 1922, 95, 165-168.

393. LUSK, G., & DUBOIS, E. F. On the constancy of the basal metabolism. *J. Physiol.*, 1924, 59, 213-216.

394. LYON, E. P., & GREISHEIMER, E. Effect of breathing dry and moist air. *Amer. J. Physiol.*, 1920, 51, 191—.

395. MACKEITH, N. W., PEMBREY, M. S., *et al.* Observations on the adjustment of the human body to muscular work. *Proc. Roy. Soc. London*, 1923-1924, 95B, 413—.

396. MAGNE, H. Les modifications des échanges respiratoires pendant l'exercise musculaire. En *Physiologie du travail.* Paris: Instit. Lannelongue d'hygiène soc., 1922. Pp. 25-68.

397. MAGNUS-LEVY, A. The physiology of metabolism. I & III. In *Metabolism and practical medicine*, Eng. ed., ed. by Van Noordon, 1907.

398. ———. Über die Grösse des respiratorischen Gaswechsels unter dem Einfluss der Nahrungsaufnahme. *Arch. f. d. ges. Physiol.*, 1894, 55, 1—.

399. MAINZER, —. Stoffwechselstudien über den Einfluss geistiger Tätigkeit und protrahierten Wachens. *Monatssch. f. Psychiat. u. Neur.*, 1903, 14, 442—.

400. MANGOLD, E. Kreislauf und Atmung. In *Körper und Arbeit*, ed. by E. Atzler. Leipzig: Thieme, 1927. Pp. 113-145.

401. MARGARIA, R. La temperatura dell' aria espirata e la ventilazione polmonare durante il bagno di mare. *Arch. di sci. biol.*, 1927, 10, 292-308.

402. MARSH, M. E. The character of energy metabolism during work. *J. Nutrition*, 1928, 1, 57-89.

403. ———. Excess metabolism and muscular efficiency for moderate work. *Amer. J. Physiol.*, 1927, 81, 497—.

404. MARTIN, E. G. Fatigue and rest. *Indus. Psychol.*, 1926, 1, 345—.

405. ———. Tests of muscular efficiency. *Physiol. Rev.*, 1921, 1, 454-475.

406. MARTIN, E. G., & GRUBER, C. M. On the influence of muscular exercise on the activity of bulbar centers. *Amer. J. Physiol.*, 1913, 32, 315—.

407. MARTINI, P., & PIERACH, A. Eine maschinelle Schwenkvorrichtung als Ergänzung zum Apparat nach Haldane-Orsat. *Arch. f. exper. Path. u. Pharm.*, 1927, 120, 341—.

408. Mason, E. H., & Mason, H. H. Effect of ultraviolet light on oxygen consumption and on total metabolism. *Arch. Internal Med.*, 1927, 39, 317-329.

409. Mathieu, P., & Schaeffer, Y. De la ventilation pulmonaire apparente: ses relations avec la fréquence respiratoire. *C. r. Soc. de biol.*, 1926, 94, Pt. 2, 1246-1248.

410. Mayer, A., & Wurmser, R. Sur une méthode de détermination du métabolisme de base. *C. r. Soc. de biol.*, 1924, 91, Pt. 1, 449-451.

411. Mayers, M. R. Fatigue in the washing departments of the N. Y. Steam Laundries. *J. Indus. Hygiene*, 1924, 6, 270—.

412. McCann, W. S. Calorimentry in medicine. (3, Pt. 1.) 1924. Also published separately under title.

413. ———. The effect of the ingestion of foodstuffs on the respiratory exchange in pulmonary tuberculosis. *Arch. Internal Med.*, 1921, 28, 847—.

414. McClendon, J. F. An improved portable calorimeter. *Proc. Soc. Exper. Biol. & Med.*, 1925, 23, 234—.

415. McClendon, J. F., Anderson, H. C., Steggerda, F. R., Conklin, C., & Whitaker, M. Simple respiration apparatus for determination of oxygen and carbon dioxide in indirect calorimetry. *J. Biol. Chem.*, 1928, 77, 413—.

416. McConnell, W. J., & Yagloglou, C. P. Basal metabolism as affected by atmospheric conditions. *Arch. Internal Med.*, 1925, 36, 382-396.

417. McConnell, W. J., Yagloglou, C. P., & Fulton, W. B. Basal metabolism before and after exposure to high temperature and various humidities. *J. Amer. Soc. Heat. & Vent. Eng.*, 1925, 31, 23-34. Also in *Pub. Health Rep.*, 1924, 39, 3075-3088.

418. McCurdy, J. F. The physiology of exercise. Philadelphia: Lea & Febiger, 1924. Pp. 237.

419. McDowall, R. J. S. The physiological effects of exercise and mental effort on the circulation and their relation to industry. *J. Indus. Hygiene*, 1928, 10, 94—.

420. McDowall, R. J. S., & Wells, H. M. Physiology of monotony. *Brit. Med. J.*, 1927, i, 414—.

421. McKinlay, C. A. Suggested adaptation to the gasometer method of the determination of the basal metabolic rate

from carbon dioxide elimination. *J. Lab. & Clin. Med.*, 1921-1922, 7, 665—.

MEANS, J. H. Dyspnoea. *Medicine*, 1924, 3, 309—.

MELKA, J. [Simple apparatus to analyze respiratory gases.] *Casopis lékaruv českych. Orgán českého lékarstva*, Praha, 1927, 66, 894—.

MELLI, G. La tecnica della determinazione del ricambio respiratorio. Un nuovo semplice apparecchio. *Arch. di pat. e clin. med.*, 1927, 6, 119-136.

MESERVE, J. C. Metabolism apparatus and technique. *Hosp. Corps Quar.*, 1927, 11, 163-168.

MESSERLE, N. Die Veränderungen in Elektrokardiogramm bei Körperarbeit. *Zsch. f. d. ges. exper. Med.*, 1928, 60, 490-501.

MOBITZ, W. (Freiburg). Über dei Korrelation von Atmung, Kreislauf und Sauerstoffverbrauch während der körperlichen Arbeit. *Klin. Woch.*, 1928, 7, 438-441.

MØLLER, E. Clinical investigations into the basal metabolism in diseases of the thyroid gland. *Acta Med. Scand., Suppl.*, 1927, No. 21. Pp. 219.

MONTORO, O. El metabolismo basal en el hombre tropical (Algunas notas aclaratorias). *Arch. de la Soc. estud. clin.*, 1927, 27, 21-49.

MOOG, O. & SCHWEIDER, —. Über den unmerklichen Gewichtsverlust bei Wettläufern und Wettschwimmern. *Klin. Woch.*, 1926, 5, 2252-2254.

MOORE, J. W. Combination mouth and nose piece designed for use with metabolic apparatus. *J. Amer. Med. Asso.*, 1921, 76, 1163—.

MORGAN, J. J. B. The overcoming of distraction and other resistances. *Arch. Psychol.*, 1916, 35. Pp. 84.

MORGULIS, S. Fasting and undernutrition—a biological and sociological study of inanition. New York: Dutton, 1927. Pp. 407.

————. Muscular work and the respiratory quotient. *Biochem. Bull.*, 1913-1914, 3, 435—.

MORSE, W. Applied biochemistry. Philadelphia: Saunders, 1927.

MOSS, K. N. Some effects of high air temperatures and muscular exertion upon colliers. *Proc. Roy. Soc. London*, 1923, 95B, 181—.

437. MOTTRAM, V. H. Physiology. New York: Norton, 1928.

438. MOWBRY, H. W. Harmful noises and their elimination. *Indus. Psychol.*, 1926, 1, 338—.

439. MÜLLER, E. A. Über den gegenwärtigen Stand der Benedict'schen Methode zur Bestimmung des Gasstoffwechsels. *Münch. med. Woch.*, 1926, 73, 1582-1583.

440. MURLIN, J. R. Normal processes of energy metabolism. *Endocrinol. & Metab.*, 1922, 3, 515—.

441. ———. On the law of surface area in energy metabolism. *Science*, 1921.

442. ———. Physiology of metabolism in infancy and childhood. *Abt's Pediatrics*, 1923, 1, 520—.

443. MUSCIO, B. Is a fatigue test possible? *Brit. J. Psychol.*, 1921, 12, 31—.

444. MYERS, C. S. Present day applications of psychology with special reference to industry, education and nervous breakdown. London: Univ. London Press, 1918. (New ed.: Mind and work; the psychological factors in industry and commerce, 1920.)

445. NETSCHAJEFF, A. Zur Frage über Ermüdungsmessungen. *Arch. f. d. ges. Psychol.*, 1928, 63, 237-248.

446. NEWCOMB, P. B. An adapted mask for basal metabolism apparatus. *J. Lab. & Clin. Med.*, 1921-1922, 7, 560-562.

447. NEWCOMER, H. S. Aids to basal metabolic rate determination. *Arch. Internal Med.*, 1922, 29, 748-762.

448. ———. A simple laboratory gas meter and an improved Haldane gas analysis apparatus. *J. Biol. Chem.*, 1921, 47, 490—.

449. NICLOUX, M. Dosage de l'acide carbonique. Applications. *C. r. Soc. de biol.*, 1927, 97, 811—.

450. NOYONS, A. K. The differential calorimeter and the determination of human basal metabolism. Louvain: René Fonteyn; New York: Stechert, 1927. Pp. 189.

451. OCHSNER, E. H. Chronic fatigue intoxication, a heretofore inadequately described affection. New York: Stechert, 1923. Pp. 143.

452. ORR, J. B., & KINLOCH, J. P. Note on the influence of diet on the energy expenditure in work. *Roy. Army Med. Corps J.*, 1921, 36, 81—.

453. ———. On the estimation of the physiological cost of muscular work. *Brit. Med. J.*, 1921, II, 39—.

454. OWEN, R. B., COPE, H. E., & HILL, L. R. Unsatisfactory basal metabolic test as a result of perforated ear drums. *J. Amer. Med. Asso.*, 1928, 90, 1442—.

455. OZORIO DE ALMEIDA, A. Le métabolisme basal de l'homme tropical. *J. de physiol.*, 1924, 22, 12-18.

456. PADGET, P. Respiratory response to carbon dioxide. *Amer. J. Physiol.*, 1928, 83, 384—.

457. PAECHTNER, J. Ein Beitrag zur Technik des Respirationsversuchs nach J. Geppert und N. Zuntz. *Biochem. Zsch.*, 1927, 186, 264-268.

458. PAGE, R. M. On supplanting the industrial fatigue concept. *J. Bus.*, 1929, 2, 137-150. Also in *Soc. Ind. Engin. Bull.* 1929, 11, No. 7, 3-9.

459. PALMER, P. E., & WEAVER, E. R. Thermal conductivity method for the analysis of gases. *Tech. Papers Bur. Stand., U. S. Dept. of Commerce*, 1924, No. 249.

460. PEABODY, F. W., & STURGIS, C. C. Clinical studies of the respiration: VII. The effect of general weakness and fatigue on the vital capacity of the lungs. *Arch. Internal Med.*, 1921, 28, 501—.

461. ———. The effect of exercise on the metabolism, heart rate and pulmonary ventilation of normal subjects and patients with heart disease. *Arch. Internal Med.*, 1922, 29, 277—.

462. PEMBERTON, R. A summary of the effects of external heat upon the human body. *Amer. J. Med. Sci.*, 1925, 169, 485—.

463. PEREIRA, J. R. A note on the technique of determining the resting oxygen intake while breathing concentrated oxygen mixtures. *Proc. Roy. Soc. London*, 1925-1926, 98B, 480-484.

464. PERWITZSCHKY, R. Die Temperatur und Feuchtigkeitsverhältnisse der Atemluft in den Luftwegen; Untersuchungen unter normalen Aussenbedingungen und bei normaler Atemgrösse am ruhenden Menschen. *Arch. f. Ohrenk.*, 1927, 117, 1-36.

465. PICKWORTH, F. A. Basal metabolism as determined by the respiratory exchange. *Proc. Roy. Soc. London*, 1927, 101B, 163—.

466. PIERCE, H. F. An integrating device for closed circuit respiration apparatus. *Lancet*, 1927, 213, Pt. 1, 340—.

467. Poffenberger, A. T. Applied psychology, its principles and methods. New York: Appleton, 1927. Pp. 586.

468. Polakov, W. N. Carbon dioxide as an index of fatigue. *Mech. Eng.*, 1925, 47, Pt. 2, 1043-1046.

469. ———. New method of indicating fatigue by measuring CO_2 in exhaled breath. *Manag. & Admin.*, 1925, 10, 343—.

470. Ponzo, M. Der Ausdruck des Willensfaktors in der Atmungskurve. *Proc. 8th Int. Cong. Psychol.*, 1927, 368-371.

471. ———. Il fattore volitivo nelle sue manifestazione respiratorie. *Arch. ital. di psicol.*, 1927, **5**, 188-198.

472. Porges, O. Über die Bedeutung der Bestimmung der CO_2-Spannung in der Alveolarluft für die Klinik. *Klin. Woch.*, 1924, 3, 209-213.

473. Poyer, G. Activité mentale, travail intellectuel et fatigue. *J. de psychol.*, 1921, **18**, 300-319.

474. Pryde, J. Recent advances in biochemistry. Philadelphia: Blakiston, 1926.

475. Rabinowitch, I. M., & Bazin, E. V. A simplified and accurate method of determining basal metabolic rates; an electrometric (katharometer) procedure. *Canadian Med. Asso. J.*, 1926, 16, Pt. 1, 638—.

476. Raper, H. S. A modified valve attachment for the Douglas respiration apparatus. *J. Physiol.*, 1919-1920, 53, cxl—.

477. Rapport, D., & Ralli, E. P. The nature of the foodstuffs oxidized to provide energy in muscular exercise. *Amer. J. Physiol.*, 1928, **83**, 450—.

478. Reach, F. Untersuchungen über die Arbeitsleistung des Menschen mit besonderer Rücksichtnahme auf ihre praktische Beziehung zum Betrieb landwirtschaftlicher Maschinen. *Landwirtsch. Jahrb., Zsch. f. wiss. Landwirtsch.*, 1908, 1053-1101.

479. Read, J. M. Correlation of basal metabolic rate with pulse rate and pulse pressure. *J. Amer. Med. Asso.*, 1922, 78, 1887-1889.

480. Regelsberger, H. Apparat zur selbsttätigen fortlaufenden Messung der alveolaren Kohlensäure. *Zsch. f. d. ges. exper. Med.*, 1928, 61, 747-756.

481. Reynolds, M. S., Severinghaus, E. L., & Stark, M. E. (See Severinghaus, Reynolds & Stark, 514.)

482. RICHE, J. A., & SODERSTROM, G. F. The respiration calorimeter of the Russell Sage Institute of Pathology in Bellevue Hospital. *Arch. Internal Med.*, 1915, 15, 805—.

483. RIESSER, O. Zur Physiologie der körperlichen Leistung. *Zsch. f. wiss. Bäderk.*, 1927, 2, 109—.

484. ROSENHEIM, O. A preliminary study of the energy expenditure and food requirements of women workers. *Proc. Roy. Soc. London*, 1919, 91B, 44—.

485. ROTH, P. Modifications of apparatus and improved technic adaptable to the Benedict type of respiration apparatus. *Bull. Battle Creek Sanatarium*, 1923, 18, 131—, Papers 3 & 4 & Appendix.

486. ———. Modifications of apparatus and improved technic adaptable to the Benedict type of respiration aparatus. Part One: Valves *vs.* the electric impeller. *Boston Med. & Surg. J.*, 1922, 186, 457-463.

487. ———. Modifications of apparatus and improved technic adaptable to the Benedict type of respiration apparatus. Part Two: Technic. *Boston Med. & Surg. J.*, 1922, 186, 463-465.

488. ———. Modifications of apparatus and improved technic adaptable to the Benedict type of respiration apparatus. Part Three: Graphic method for the estimation of the metabolic rate. *Boston Med. & Surg. J.*, 1922, 186, 491-498.

489. ———. Modifications of apparatus and improved technic adaptable to the Benedict type of respiration apparatus. Part Four: Moisture absorbing efficiency of CO_2 absorbents. *Boston Med. & Surg. J.*, 1922, 186, 498-501.

490. ROWE, A. W. New mask for use in basal metabolism determinations. *J. Lab. & Clin. Med.*, 1926-1927, 12, Pt. 1, 590—.

491. ROWE, A. W., & EAKIN, M. Monthly fluctuations in the normal metabolic rates of men and women. *Calif. State Med. J.*, 1921, 19, 320—.

492. Royal Society Food (War) Committee. Report on the food requirements of man. *Roy. Soc. London*, 1919.

493. RYAN, A. H. The quantitative measurement of general fatigue. *Amer. J. Physiol.*, 1918, 45, 537—.

494. SACHSENBERG, E. Ergebnisse wissenschaftlicher Forschung auf dem Gebiete der Ermüdung und ihre Anwendbarkeit

auf die Praxis. (*Zentbl. f. Gewerbehygiene u. Unfallverhütung,* Beiheft 7, Arbeit und Ermüdung.) Berlin: Springer, 1927. Pp. 46-65.

495. SAINTON, P., & PÉRON, N. De l'importance relative des divers facteurs conditionnant les variations du métabolisme basal. *Paris méd.,* 1923, 49, 465-467.

496. SANBORN, F. B. [Ed.] Basal metabolism, its determination and application. Chicago: Sanborn, 1922. Pp. 282.

497. SARGENT, R. M. Recovery from vigorous exercise of short duration. *Proc. Roy. Soc. London,* 1926, 100B, 440—.

498. SCHADOW, H. Ein vereinfachter Respirationsapparat. *Klin. Woch.,* 1925, 4, 1548—.

499. SCHNEIDER, A. Respiration test as a possible substitute for the usual basal rate determinations. *J. Amer. Pharm. Asso.,* 1927, 16, 616-623.

500. SCHNEIDER, E. C. A cardiovascular rating as a measure of physical fatigue and efficiency. *J. Amer. Med. Asso.,* 1920, 74, 1507—.

501. ————. Physiological observations following descent from Pike's Peak to Colorado Springs. *Amer. J. Physiol.,* 1913, 32, 295—.

502. SCHNEIDER, E. C., & CLARK, R. W. Studies on muscular exercise under low barometric pressure: I. The consumption of oxygen and the oxygen debt. *Amer. J. Physiol.,* 1925, 74, 334-353.

503. ————. Studies of muscular exercise under low barometric pressure: II. The frequency and volume of respiration. *Amer. J. Physiol.,* 1925-1926, 75, 297-307.

504. ————. Studies of muscular exercise under low barometric pressure; output of carbon dioxide. *Amer. J. Physiol.,* 1928, 85, 65-77.

505. SCHNEIDER, E. C., CLARK, R. W., & RING, G. C. The influence of physical training on the basal respiratory exchange, pulse rate and arterial blood pressure. *Amer. J. Physiol.,* 1927, 81, 255—.

506. SCHNEIDER, E. C., & TRUESDELL, D. The effects on the circulation and respiration of an increase in CO_2 content of the blood in man. *Amer. J. Physiol.,* 1922, 63, No. 1.

507. SCHNEIDER, E. C., TRUESDELL, D., & CLARK, R. W. Oxygen consumption in men during short exposures to low barometric pressures. *Amer. J. Physiol.,* 1924, 70, 283-299.

SCHROETTER, —. Zur Kenntnis des Energieverbrauches beim Maschinenschreiben. Wien: 1925.

SCOTT, R. W. The significance of undissociated carbon dioxide in respiration. *Amer. J. Physiol.*, 1918, 47, 43—.

SCOTT, V. T. The application of certain physical efficiency tests. *J. Amer. Med. Asso.*, 1921, 76, 705—.

SEGAL, H. L., BINSWANGER, F. G., & STROUSE, S. The effect of emotion on the basal metabolism. *Proc. Soc. Exper. Biol. & Med.*, 1927, 24, 845-846. Also in *Arch. Internal Med.*, 1928, 41, 834-842.

SEUFFERT, R. W., & NITSCHKE, M. Ueber einige Kontrollversuche an einem kleinen Respirationsapparat nach Voit. *Beitr. z. Physiol.*, 1925, 3, 61-78.

SEVERINGHAUS, E. L. A simple bicycle ergometer. *Amer. J. Physiol.*, 1927, 80, 347—.

SEVERINGHAUS, E. L., REYNOLDS, M. S., & STARK, M. E. Human energy metabolism: II. The mechanical efficiency of the body on carbohydrate, fat and mixed diets. *Amer. J. Physiol.*, 1927, 80, 355-362.

SHEPARD, G. A. Fatigue experiments at Purdue University. *Indus. Manag.*, 1921, 62, 281—; 63, 354—.

SIMONSON, E. Ermüdung und Erholung nach körperlicher Arbeit. *Klin. Woch.*, 1926, 5, 2337-2340.

————. Ein neuer Respirationsapparat. *Arbeitsphysiol.*, 1928, 1, 224-257.

————. Die Wirkung verstärkter willkürlicher Atmung auf die Geschwindigkeit der Erholung nach körperlicher Arbeit. *Arbeitsphysiol.*, 1928, 1, 87-102.

————. Zur Physiologie des Energieumsatzes beim Menschen, I. Beiträge zur Physiologie der Arbeit, der Restitution und der Atmung. *Arch. f. d. ges. Physiol.*, 1926, 214, 380-402.

————. Zur Physiologie des Energieumsatzes beim Menschen, II. Zur Physiologie des Stehens. *Arch. f. d. ges. Physiol.*, 1926, 214, 403-415.

————. Zur Physiologie des Energieumsatzes beim Menschen, III. Weitere Beiträge zur Physiologie der Erholung bei körperlicher Arbeit. *Arch. f. d. ges. Physiol.*, 1927, 215, 716-742.

————. Zur Physiologie des Energieumsatzes beim Men-

schen, V. Weitere Beiträge zur Physiologie der Atmung und der Übung. *Arch. f. d. ges. Physiol.*, 1927, 215, 752-767.

523. SIMONSON, E., & REISSER, O. Zur Physiologie des Energieumsatzes beim Menschen, IV. Zur Physiologie der Übung. *Arch. f. d. ges. Physiol.*, 1927, 215, 743-751.

524. SMART, W. A. M. A simple slide rule for the calculation of respiratory quotients. *J. Physiol.*, 1927-1928, 64, ix—.

525. SMITH, H. M. Gaseous exchange and physiological requirements for level and grade walking. *Carnegie Instit. Wash. Publ.*, 1922.

526. SMITH, J. H. Basal metabolism; influence of work with special reference to thyroid gland. *Arch. Internal Med.*, 1928, 42, 47-52.

527. SMITH, J. H., & SMITH, M. C. Calculation in basal metabolism determinations. *Boston Med. & Surg. J.*, 1922, 186, 641-642.

528. SNOOK, H. C. Progress report of Committee on Elimination of Harmful Noise. *16 Ann. Safety Cong., Nat. Safety Coun.*, 1928.

529. SODERSTROM, G. F., BARR, D. P., & DUBOIS, E. F. The effect of a small breakfast on heat production. *Arch. Internal Med.*, 1918, 21, 613—.

530. SODERSTROM, G. F., MEYER, A. L., & DUBOIS, E. F. Clinical calorimetry. XI. A comparison of the metabolism of men flat in bed and sitting in a steamer chair. *Arch. Internal Med.*, 1916, 17, 872-886.

531. SOTO, M., & TORINO, A. Appareil simplifié pour déterminer le métabolisme basal. *C. r. Soc. de biol.*, 1926, 94, Pt. 1, 473-475.

532. SPAETH, R. A. The prevention of fatigue in manufacturing industries. *J. Indus. Hygiene*, 1919, 1, 435—.

533. ———. The problem of fatigue. *J. Indus. Hygiene*, 1919, 1, 22—.

534. SPOONER, H. J. Health problems involved in noise and fatigue. *Nation's Health*, 1922, 4, 91—.

535. ———. Progress of the movement for the reduction of noise. *Soc. Indus. Eng.*, 1922, 180—.

536. ———. The year's progress in the reduction of noise. *Nation's Health*, 1922, 4, 368—.

STARR, H. E. Psychological concomitants of high alveolar carbon dioxide, a psycho-biochemical study of the etiology of stammering. *Psychol. Clin.*, 1928, 17, 1-12.

STEGGERDA, M., & BENEDICT, F. G. Basal metabolism of some browns and blacks in Jamaica. *Amer. J. Physiol.*, 1928, **85**, 621-633.

STEWART, G. N. A manual of physiology with practical exercises. (11th ed.) New York: Wood, 1921. Pp. 1245.

STILES, P. G. Nutritional physiology. (5th ed.) Philadelphia: Saunders, 1924.

STOLZ, E. Ein Beitrag zur Methodik der Grundumsatzbestimmungen mit dem Kroghschen Apparat. *Wien. klin. Woch.*, 1926, 39, 1018—.

STONER, W. H. The organization of a metabolism laboratory. *J. Lab. & Clin. Med.*, 1925-1926, 11, 113-115.

————. Notes on basal metabolism: VIII. Tables of values of the DuBois surface area formula. *J. Lab. & Clin. Med.*, 1925-1926, **11**, 355-368.

————. Notes on basal metabolism: IX. Simplified calculation for gasometer gas analysis method. *J. Lab. & Clin. Med.*, 1927, 12, 884-888.

————. Notes on basal metabolism: X. Simplified data blank for gasometer gas analysis method. *J. Lab. & Clin. Med.*, 1927, 13, 164—.

————. Notes on basal metabolism: XI. Basal metabolism standards. *J. Lab. & Clin. Med.*, 1927, 13, 265-269.

STRAUSS, W. Über eine neue Methode zur Messung körperlicher Ermüdung. *Klin. Woch.*, 1927, 6, 483—.

STRAUSS, W., & BANDMANN, M. Probleme und Methoden der Ermüdungsmessung. *Zentbl. f. d. ges. Hygiene*, 1927, 15, 753-774.

STUDER, F. Der Sauerstoffverbrauch beim Gehen auf horizontaler Bahn. *Arch. f. d. ges. Physiol.*, 1926, **212**, 105-120.

SÜFFLE, K., HOFFMAN, P., & WALZ, L. Eine vereinfachte Methode zur quantitativen Bestimmung von Kohlensäure, Ammoniak und Schwefelwasserstoff in der Luft bewohnter Räume. *Arch. f. Hygiene*, 1927, 98, 147—.

SUMNER, J. B. Textbook of physiological chemistry. New York: Macmillan, 1927.

551. SUNDSTROEM, E. S. Physiological effects of tropical climate. *Physiol. Rev.*, 1927, 7, 320—.

552. Symposium. Studies in industrial psychology. *Int. Labour Rev.*, 1922, 6, 988—.

553. TAKAHIRA, H. Report of metabolic laboratory. *Imperial Govt., Instit. Nutrition*, Tokyo, 1925, 1, No. 1.

554. TAKAHIRA, H., & ISHIBASHI, H. [On the influence of psychical work on energy metabolism.] *4th Gen. Meet. Nat. Physiol. Soc.*, Moukden, Aug. 22-23, 1925.

555. TALBOT, F. B. Basal metabolism of children. *Physiol. Rev.*, 1925, 5, 477—. Also in *Monatssch. f. Kinderhk.*, 1924, 27, 465—.

556. TALBOT, J. H., FÖLLING, A., HENDERSON, L. J., DILL, D. B., EDWARDS, H. T., & BERGGREN, R. E. L. Studies in muscular activity: V. Changes and adaptations in running. *J. Biol. Chem.*, 1928, 78, 445-463.

557. TAYLOR, F. W. The principles of scientific management. New York: Harper, 1913.

558. TERROINE, E. F., & ZUNZ, E. Le métabolisme de base. Physiologie, pathologie (les problèmes biologiques). Paris: Presse Univ. de France, 1928. Pp. 187.

559. THANNHAUSER, S. J. Die Probleme des Stoffwechsels im Wandel der Zeiten. *Münch. med. Woch.*, 1928, 75, 76-79.

560. TIGERSTEDT, R. Die Physiologie des Stoffwechsels. *Nagel's Handb. d. Physiol. d. Menschen*, 1905, 1, 331-557.

561. ———. Respirationsapparate. *Tigerstedt's Handb. d. physiol. Meth.*, 1911, 1, Pt. 3.

562. TISSOT, J. Nouvelle méthode de mesure et d'inscription du débit et des mouvements respiratoires de l'homme et des animaux. *J. de physiol.*, 1904, 6, 688-700.

563. TOTTEN, E. Oxygen consumption during emotional stimulation. *Comp. Psychol. Monog.*, 1925, 3, No. 13. Pp. 79.

564. TURNER, A. H. The circulatory minute volumes of healthy young women in reclining, sitting, and standing positions. *Amer. J. Physiol.*, 1927, 80, 601—.

565. TURNER, E. L. The racial factor in basal metabolism. *J. Amer. Med. Asso.*, 1926, 87, 2052—.

566. VAN SLYKE, D. D. The relation of carbon dioxide and oxygen. *Trans. Asso. Amer. Physicians*, 1924, 39, 57—.

567. ———. The carbon dioxide carriers of the blood. *Physiol. Rev.*, 1921, 1, 141—.
568. VERNON, H. M. Industrial fatigue in relation to atmospheric conditions. *Physiol. Rev.*, 1928, 8, 130-150.
569. ———. Industrial fatigue and efficiency. New York: Dutton, 1921.
570. VERNON, H. M., *et al.* Methods of investigating ventilation and its effects. (*Spec. Rep., Med. Res. Coun., No. 100.*) London: H. M. Stationery Office.
571. VIALE, G. Il consumo energetico nel lavoro umano in varie condizioni sperimentali. Nota V. Azione del clima alpino. *Arch. di sci. biol.*, 1924, 6, 350—.
572. ———. Il consumo energetico nel lavoro umano in varie condizioni sperimentali. Nota VI. Azione della temperatura sul consumo energetico dopo il lavoro. *Arch. di sci. biol., 1925, 7, 363—.*
573. ———. Il consumo energetico nel lavoro umano in varie condizioni sperimentali. Nota VII. Il laboro eseguito dai mancini dagli ambidestri e dai destri. *Arch. di sci. biol.*, 1925, 7, 373—.
574. ———. Il consumo energetico nel lavoro umano in varie condizioni sperimentali. Nota VIII. Il lavoro nel digiuno e dopo i pasti. *Arch. di sci. biol.*, 1927, 10, 401-407.
575. ———. Il consumo energetico nel lavoro umano in varie condizioni sperimentali. Nota IX. Eliminazione di anidride carbonica nello sforzo massimo. Determinazione di una costante individuale. *Arch. di sci. biol.*, 1928, 11, 63—.
576. ———. El consumo energético del trabajo humano en varias condiciones experimentales. Nota VIII. El trabajo en ayunas y después de las comidas. *Rev. de Soc. Argent. de biol.*, 1927, 3, 419-424.
577. ———. El consumo energético del trabajo humano en varias condiciones experimentales. Nota IX. Eliminación de anhídrido carbónico en el esfuerzo máximo. Determinación de una constante individual. *Rev. de Soc. Argent. de biol.*, 1927, 3, 427-436.
578. Vienna letter. Modern conceptions of the physiology of work. *J. Amer. Med. Asso.*, 1926, 86, 363—.
579. VOLSHINSKI, V. A., & YAKOVENKO, V. A. [The productiveness of man's work in high atmospheric temperatures,

ascertained by the method of exchange of gases.] [*J. Exper. Biol. & Med.*], 1927, No. 19, 209-223.

580. WALLER, A. D. Introduction to human physiology. (3rd ed.) 1896.

581. ———. Calibration of a dock laborer by means of his CO_2 discharge. *J. Physiol.*, 1918-1919, 52, xlviii—.

582. ———. The cost of mechanical work in terms of CO_2 expired. *J. Physiol.*, 1918-1919, 52, l-li.

583. ———. The CO_2 ordinate of a dock laborer during six days' work. *J. Physiol.*, 1918-1919, 52, lix—.

584. ———. The time relation between muscular work and CO_2 discharge. *J. Physiol.*, 1918-1919, 52, lxvii-lxviii.

585. ———. 'Work' measured by CO_2. *J. Physiol.*, 1918-1919, 52, lxxii-lxxiii.

586. ———. The physiological cost of marching measured by CO_2. *J. Physiol.*, 1919-1920, 53, xxiv—.

587. WALLER, A. D., & DEDECKER, G. An examination of the mechanical efficiency of a healthy adult by his CO_2 discharge. *J. Physiol.*, 1919-1920, 53, xxx.

588. ———. Bicycle as compared with staircase ergometry. *J. Physiol.*, 1919-1920, 53, xliv.

589. ———. The physiological cost of collier's work. *J. Physiol.*, 1920-1921, 54, lxii.

590. ———. The physiological cost of a tailor's work measured by CO_2 and expressed in calories. *J. Physiol.*, 1919-1920, 53, lxxiii.

591. ———. The physiological cost of walking; in and out of training. *J. Physiol.*, 1920-1921, 54, lxxxi.

592. ———. The physiological cost of printer's work measured by CO_2 and expressed in calories. *J. Physiol.*, 1919-1920, 53, civ.

593. ———. The physiological cost of work in various departments of the Times Printing House. *J. Physiol.*, 1919-1920, 53, cv.

594. WALLER, A. D. The physiological cost of muscular work. *Brit. Med. J.*, 1920, Apr. 17.

595. ———. The physiological cost of muscular work. *Brit. Med. J.*, 1921, Pt. 2, 627—.

596. WALLER, A. D., & DEDECKER, G. The physiological cost

of muscular work measured by the exhalation of carbon dioxide. *Brit. Med. J.*, 1921, May 7, 669—.

597. ———. La dépense physiologique (exhalation de CO_2) dans la marche sur tapis boulant et sur terre ferme. *C. r. Soc. de biol.*, 1921, 84, 910-912.

598. ———. Observation comparative de la dépense physiologique de la marche, exprimée en calories: A, d'après le CO_2 et O_2; B, d'apres le CO_2 seul. *C. r. Soc. de biol.*, 1921, 85, 853—.

599. ———. Dépense en CO_2 pendant la nage. *C. r. Soc. de biol.*, 1921, 85, 902-903.

600. WALLER, A. D. La dépense physiologique, mesurée chez le soldat en marche de route. *Bull. Acad. roy. de méd. de Belg., Brux.*, 1920, 4th ser., 30, 1049-1056.

601. ———. Measurement of the energy output of 'heavy workers.' *Rep. Brit. Asso. Adv. Sci.*, 1919, 310—.

602. ———. The energy output of dock laborers during 'heavy work.' *Proc. Roy. Soc. London*, 1919, 91B, 166—.

603. WALLER, A. D., & DEDECKER, G. The energy output of laborers on cold storage work. *Proc. Roy. Soc. London*, 1919, 91B, 229—.

604. WALLER, AUGUSTUS DÉSIRÉ, (obituary). *Proc. Roy. Soc. London*, 1922, 93B, xxvi.

605. WALTHER, L. La technopsychologie du travail. Collection d'actualités pédagogiques. Neuchâtel-Paris: Delachaux & Niestlé, 1926. Pp. 239.

606. WANG, C. C., & KERN, R. Influence of sleep on basal metabolism in children. *Proc. Instit. Med.*, 1927, 6. Pp. 253.

607. WANG, C. C., STROUSE, S., & SMITH, E. A. Influence of fatigue on the heat production during muscular work in obese, normal, and thin subjects. *J. Biol. Chem.*, 1927, 74, xxxvii-xxxix.

607*a*. WARDLAW, H. S. H. Gravimetric determination of basal metabolism. *Med. J. Australia*, 1927, 1, 506—.

608. WASSERMANN, M. Die physiologischen Grundlagen des Kohlensäurebestandes. *Wien. med. Woch.*, 1925, 75, 1663—.

609. WEBER, W. Praktische Psychologie im Wirtschaftsleben. Leipzig: Barth, 1927. Pp. 418.

610. WENZIG, K. Beiträge zur Physiologie des Schaufelns. *Arbeitsphysiol.*, 1928, 1, 154-187.

611. WHEELER, W. H., JR. Measuring the energy cost of work. *Indus. Psychol.*, 1926, 1, 627—.

612. WILLARD, H. H., & SMITH, G. F. The preparation and properties of magnesium perchlorate and its use as a drying agent. *J. Amer. Chem. Soc.*, 1922, 44, 2255—.

613. WILLIAMS, C. D., & BENEDICT, F. G. Basal metabolism of Mayas in Yucatan. *Amer. J. Physiol.*, 1928, 85, 634-649.

614. WILLIAMS, J. B., & NOLTING, M. Dental contribution to basal metabolism test. *J. Lab. & Clin. Med.*, 1926, 11, 1097—.

615. WILSON, D. R. International co-operation in the study of industrial fatigue. *Int. Labor Office Stud. & Repr.*, 1923, Ser. F, No. 9, 35—.

616. ————. On some recent contributions to the study of industrial fatigue. *J. Roy. Statis. Soc.*, 1923, 86, 483—.

617. WILSON, J. R., LEVINE, S. Z., RIVKIN, H., & BERLINER, F. Respiration metabolism in infancy and in childhood. Effect of moderate exercise on respiratory quotient; study of four children. *Amer. J. Dis. Child.*, 1927, 33, 618—.

618. WILTSHIRE, M. O. Some observations on basal metabolism in menstruation. *Lancet*, 1921, 201, Pt. 1, 388—.

619. WISHART, G. M. The influence of previous muscular activity and other factors on the basal metabolism. *Quar. J. Med.*, 1927, 20, 199—.

620. ————. The influence of the protein intake on the basal metabolism. *J. Physiol.*, 1928, 65, 243-254.

621. ————. The variability of basal metabolism. *Quar. J. Med.*, 1927, 20, 193—.

622. ZEIGLER, L. H., & LEVINE, B. S. The influence of emotional reactions on basal metabolism. *Med. Press*, 1925, N. S. 119, 70-73. Also in *Amer. J. Med. Sci.*, 1925, 169, 68-76.

623. ZUNTZ, N. Stoff- und Kraftwechsel. Wärmehaushalt. Cap. 22 im *Lehrbuch der Physiologie des Menschen*, 3rd ed. Leipzig: 1920.

624. ZUNTZ, N., & SCHUMBURG, —. Studien zu einer Physiologie des Marsches. Berlin: 1901.

625. ZURAWLEW, I. N., & FELDMANN, A. B. Über den Einfluss der Blutstromunterbrechung auf die Restitution des ermüdeten menschlichen Muskels. (Beiträge zur Frage über die Lokalisation der Ermüdung bei Muskelarbeit.) *Arbeitsphysiol.*, 1928, 1, 187-197.

Bibliography of Bibliographic Sources

The following is a list of bibliographic sources other than the bibliographies appended to various references in the preceding list.

626. *Abstract of the literature of industrial hygiene*, Vol. 1–, 1919–. Appears as a part of *The Journal of Industrial Hygiene*, Harvard School of Public Health.

627. *American Journal of Physiology, Index*: Vol. 1, 1898-1912, Vol. 2, 1912-1922.

628. *Arbeitsphysiologie. Zeitschrift für die Physiologie des Menschen bei Arbeit und Sport*, ed. by E. Atzler and M. Rubner, Vol. 1–, 1928-1929–. Berlin: Springer.

629. Beresford, Adams-Clark. Industrial psychology bibliography. *British Library of Political Science Bulletin*, 1924.

630. Berg, R. M. Bibliography of management literature. New York: American Society of Mechanical Engineers, 1927. Pp. 67.

631. *Berichte über die gesamte Physiologie und experimentelle Pharmacologie*, Vol. 1–, 1920–. Berlin: Springer.

632. Bibilographie d'hygiène industrielle; Bibliography of industrial hygiene; and Literatur der Gewerbehygiene. Geneva, Switzerland: International Labor Office.

633. Bibliography of bibliographies on industrial engineering and management. The Society of Industrial Engineers, 1920. Pp. 23.

634. *Bulletin of the Public Affairs Information Service*, Vol. 1–, 1915–. New York: Public Affairs Information Service.

635. Cannons, H. G. T. Bibliography of industrial efficiency and factory management. London: Routledge, 1920.

636. Index Catalogue of the Library of the Surgeon-General's Office, U. S. Army: First Series, 1880-1895; Second Series, 1896-1914; Third Series, 1918—.

637. *Index Medicus*: First Series, monthly. Published privately, 1879-1899.

638. *Index Medicus*: Second Series, monthly. Carnegie Institute of Washington, 1903-1920.

639. *Index Medicus*: Third Series, quarterly. Carnegie Institute of Washington, 1921-1926.

640. *Jahresbericht über die gesamte Physiologie und experimentelle Pharmakologie mit vollständiger Bibliographie*, annual, Vol. 1–, 1920–. Berlin: Springer.

641. *Journal of the National Institute of Industrial Psychology*, Vol. 1–, 1922-1923–. London: National Institute of Industrial Psychology.

642. *Physiological Abstracts*, monthly, Vol. 1–, 1916-1917–. Physiological Society (Great Britain and Ireland).

643. *Physiological Reviews*, quarterly, Vol. 1–, 1921–. Baltimore, Md.: American Physiological Society.

644. *Psychological Abstracts*, monthly, Vol. 1–, 1927–. Princeton, N. J.: Psychological Review Co.

645. *Psychological Bulletin*, Vol. 1–, 1904–. Princeton, N. J.: Psychological Review Co.

646. *Psychological Index*, annual, Vol. 1–, 1894–. Princeton, N. J.: Psychological Review Co.

647. *Quarterly Cumulative Index*, 1916-1926. Chicago: American Medical Association.

648. *Quarterly Cumulative Index Medicus*, Vol. 1–, 1927–. Chicago: American Medical Association.

649. Rossi, W. H., & Rossi, D. I. P. Personnel administration, a bibliography. Baltimore, Md.: Williams and Wilkins, 1925. Pp. 365.

LA MESURE DE L'ÉNERGIE HUMAINE EMPLOYÉE DANS L'INDUSTRIE

Guide Général à ce qu'on a écrit à ce sujet

(Résumé)

Cette monographie a comme but la présentation aux lecteurs américains de ce qu'on a écrit sur la mesure de l'emploi de l'énergie. Bien que des hommes de science allemands aient fait beaucoup d'études du métabolisme du travail lesquelles possèdent un intérêt industriel, et des investigateurs français et anglais se soient servis assez bien des déterminations de l'échange respiratoire dans les études industrielles, on n'a fait de telles expériences aux Etats-Unis que d'une façon très limitée. Il n'existe aucun manuel général en anglais sur les méthodes, les appareils, et les résultats dans le champ de la mesure de l'énergie humaine employée dans le travail, de sorte que l'expérimentation de ce genre au laboratoire ou dans l'industrie est impossible pour les psychologues industriels et les ingénieurs industriels à moins qu'ils se connaissent très bien en physiologie ou qu'ils sachent lire les manuels allemands à ce sujet.

On a écrit cette monographie seulement pour servir de guide à ce qu'on a écrit à ce sujet et l'on ne présente aucuns résultats de l'expérimentation originale. Puisqu'il n'y a pas de conclusions expérimentales à résumer, on ne présente ci-dessous qu'une énumération des divers sujets et titres, lesquels montrent l'organisation et les matières discutées.

Fondations Physiologiques. Théorie de la Détermination de l'Echange respiratoire; La Fatigue et la Perte de l'Energie; la Mesure de la Fatigue; la Physiologie générale; le Métabolisme fondamental; l'Echange métabolique général et le Métabolisme de l'Energie; La Physiologie respiratoire et la Chimie du Sang dans ses rapports au Travail musculaire; la Bio-Chimie et la Dynamique de l'Action musculaire; la Physiologie du Travail musculaire; le Quotient respiratoire; la Computation des Calories; les Facteurs qui Influent sur le Métabolisme; les Facteurs physiologiques sujets au contrôle expérimental; le Travail mental; les Facteurs du Milieu qui Influent sur le Métabolisme; Références générales sur les facteurs qui influent sur le Métabolisme.

Appareils et Methodes. Manuels généraux sur les Appareils et les Méthodes; Indices fortuits et non-quantitatifs de la Vitesse du Métabolisme; la Calorimétrie directe; la Calorimétrie indirecte; les Méthodes de Circuit ouvert, type gasomètre; les Méthodes de Circuit ouvert, type Sac Douglas; Critique et Défense de la Méthode Waller; Analyse des Gaz par la Méthode Waller; Autres Méthodes Simplifiées de l'Analyse des Gaz; Autres Méthodes de Circuit ouvert de la Détermination de la Vitesse du Métabolisme; la Ventilation pulmonaire totale comme indice de l'Energie employée dans le Travail léger; Méthodes de Circuit fermé pour déterminer la consommation de l'oxygène et la quantité du gaz carbonique; Méthodes graphiques pour déterminer l'Echange respiratoire; Appareils de Respiration pour les Expériences de l'Echange respiratoire; Appareils accessoires pour les Expériences de l'Echange respiratoire.

Applications et Resultats. Les Emplois de la Détermination de l'Echange respiratoire dans l'Industrie; Études dans le laboratorie des Composants du Travail industriel; l'Energie employée dans les occupations industrielles; l'Energie employée quand on marche à pied; l'Étude des Conditions du Travail industriel; l'Effet du bruit dans le Travail industriel; Périodes de Repos; Tests métaboliques de l'Efficience physique.

PAGE

DIE MESSUNG DES MENSCHLICHEN ENERGIEVERBRAUCHS IN DER INDUSTRIE: EIN ALLGEMEINER WEGWEISER ZUR LITERATUR

(Referat)

Es ist der Ziel dieser Monographie, die Literatur über die Messung des Energieverbrauchs für Amerikanische Leser darzubieten. Obwohl deutsche Forscher viele Untersuchungen über den metabolischen Austausch während der Arbeit durchgeführt haben die von industrieller Interesse sind, und obwohl französische und englische Arbeiter einen ziemlich weiten Gebrauch von Bestimmungen des respiratorischen Haushaltes (respiratory exchange) in industriellen Untersuchungen gemacht haben, sind solche Untersuchungen in den Vereinigten Staaten nur in sehr beschrenktem Masse durchgeführt worden.

Es gibt auf Englisch kein allgemeines Handbuch über Methoden, Vorrichtungen oder Befunde im Bereich der Messung der Energiekosten der menschlichen Arbeit. Forschung dieser Art im Laboratorium oder in der Industrie bleibt industriellen Psychologen und Ingenieuren also versagt, es sei denn dass sie gründlich in der Physiologie geschult sind oder die deutschen Handbücher über diese Sache lesen können.

Die gegenwärtige Monographie soll bloss als Wegführer zur Literatur dienen, und es werden darin keine Befunde aus eigenen Versuchen gegeben. Da es keine experimentellen Befunde zu referieren gibt, wollen wir unten bloss die verschiedenen Themen und Überschrifte (topics and headings) eintragen. Hierdurch wird der Plan und der Gegenstand der Untersuchung dargelegt werden.

Physiologische Grundlagen. Theorie der Bestimmung des respiratorischen Austausches (respiratory exchange); Die Ermüdung und der Energieverbrauch; Die Messung der Ermüdung; Allgemeine Physiologie; Der Grundaustausch (basal metabolism); Allgemeiner metabolischer Austausch und Energieaustausch; Atmungsphysiologie und Blutchimie im Zusammenhang mit Muskelarbeit; Biochimie und Dynamik der Muskeltätigkeit; Physiologie der Muskelarbeit; Atmungsquotient (respiratory quotient); Berechnung der Kalorien; Einwirkungen die den Metabolismus bestimmen; Physiologische Einwirkungen die sich experimentell kontrolieren lassen; Die geistige Arbeit; Einwirkungen der Umgebung auf den Metabolismus; Allgemeine Literatur über Einwirkungen auf den Metabolismus.

Vorrichtungen (Apparatus) und Methoden. Allgemeine Handbücher über Einrichtungen und Methoden; Beiläufige und nicht-quantitative Indizien der Schnelligkeit des Metabolismus (metabolic rate); Die direkte Kalorimetrie; Indirekte Kalorimetrie; Methoden mit offenem Stromkreis, vom Typus des Douglasschen Saks (Douglas bag); Die Wallersche Vereinfachung der "Douglas bag" Methode; Kritique und Verteidigung der Methode Waller; Andere Methoden mit Gebrauch von offenem Stromkreis zur Bestimmung der Schnelligkeit des Metabolismus; Die gesammte Lungenventilation als Index des Energieverbrauchs bei leichter Arbeit; Methoden mit Gebrauch von geschlossenem Stromkreis zur Bestimmung des Sauerstoffverbrauches und des Ertrages an Kohlstoffsauere; Graphische Methoden bei der Bestimmung des Atmungsaustausches (respiratory exchange); Atmungsapparat für Versuche an dem respiratorischen Austausch; Zusatzapparat fuer Versuche an dem respiratorischen Austausch.

Anwendungen und Befunde. Nützen der Bestimmung des respiratorischen Austausches in der Industrie; Untersuchungen im Laboratorium an den Bestandteilen der industriellen Arbeit; Energieverbrauch bei Indus-

trlellen Berufen; Energieverbrauch beim Gehen; Die Untersuchung industrieller Arbeitsbedingungen; Die Einwirkung geräuschvoller Arbeitsbedingungen; Ruhepausen; Metabolische Prüfungen der körperlichen Leistungsfähigkeit.

PAGE